MODERN EDUCATION SERIES

INTRODUCTION
TO THE USE OF
STANDARDIZED TESTS

DENTON L. GEYER, Ph. D.

The PLYMOUTH PRESS
CHICAGO

MODERN EDUCATION SERIES

Edited by JAMES E. McDADE

INTRODUCTION
TO THE USE OF
STANDARDIZED TESTS

By

DENTON L. GEYER, Ph. D.

Department of Education
Chicago Normal College

THE PLYMOUTH PRESS
CHICAGO

CONTENTS

FOREWORD

This little outline is written to give the classroom teacher a brief general survey of the measuring movement in education. It attempts not so much to bring out things that are new as to set forth a certain number of the more salient facts about its subject in simple and non-technical language. It therefore touches on a considerable range of topics: how the movement has developed and what it implies; what goes into standardized tests and what they are used for; how to choose the best from among them; and how to interpret their scores.

The four or five excellent treatises we have on standardized tests are longer and cover somewhat different ground; some limit their material for the most part to descriptions of the existing tests and directions for administering them; others give some hint of the meaning of the movement as a whole, but introduce such material only incidentally and sketchily; others give their principal emphasis to teaching-devices for dealing with the defects which standarized tests may reveal; others cover a wider range of topics, but couch their ideas in highly technical terminology. Few have given full attention to both intelligence tests and achievement tests within the same covers. All use for their explanation many more pages than the average teacher feels that she could take time to make her way through.

The present brief booklet will fulfill its purpose if it gives the classroom teacher without previous experience in this field and without special psychological training a short statement of the facts about educational measurements which is at the same time readable and practically useful.

CHAPTER I

THE FUNCTIONS OF STANDARDIZED TESTS

We may think of a test as *standardized* when it has been given under uniform conditions to a sufficiently large number of children to allow us to base on these scores standards of attainment for other children of the same age or school grade. The test is thus standardized in two senses—in the sense that the method of giving and scoring it is so definitely controlled as to be standardized, and in the sense that the scores already secured serve as objective standards for other classes.

The advantages of such a test are obvious. By it the teacher can compare her class with other classes in other schools. She can find out whether her pupils are as intelligent, or as proficient in their studies, as the children of other cities or of the country in general. She can find out in which studies they are up to standard and in which below standard, and can distribute her teaching emphasis accordingly. She can discover whether her pupils have an average amount of native ability, or are decidedly inferior or superior. This helps her to decide how rapidly to proceed in her work and how much time to give to drills and reviews, as well as letting her know whether the results she is getting are of the kind that should be expected.

By an objective test the teacher can also discover whether a certain method of teaching is proving effective. She can learn this by giving at the beginning of the experiment a test of intelligence and a test of achievement in the given study, and another test of achievement in the study at the end of the experiment. The intelligence test will show whether the pupils are of a type to make slow, average, or rapid progress. The difference between the

scores on the first and second achievement tests will show how much progress, with her help, they have actually made in this study. If it is less than it should be, other teaching methods may be tried; if it is more than most children of this ability make, the teaching-method may be retained and tried again, and if again vindicated may finally be accepted with complete confidence as a good scheme for teaching that subject. This experiment cannot be carried on by means of the old-fashioned tests, because the teacher can never be sure that the two tests used at the beginning and end of the experiment are of exactly equal difficulty, or that she has graded both tests of papers with exactly the same degree of strictness. Standardized tests are so constructed that answers are either right or wrong: they will therefore be scored in just the same way at all times. And the two tests of such a pair as would be used at the beginning and end of this experiment are known to be of equal difficulty because they have been tried out with hundreds of children before being published.

To be able to see whether pupils are up to standard in each of their studies, to be able to determine their brightness as compared with other children, and to be able to try out teaching plans by the method of scientific experiment, are tremendous advantages. In these ways the teacher can learn what kind of material she has to work with, how well it has so far been worked with by herself and her predecessors, and to what extent her pet schemes of teaching are bringing results. These three uses of standardized tests, if actually in effect, would completely transform the average school. They are, however, by no means the only uses to which such tests can be put.

Standardized tests assist one in learning not only how rapidly the pupils are progressing, but why certain of them do not make progress as rapidly as they should. In other words, standardized tests can be used for *diagnosis* of

mental difficulties, just as a clinical thermometer can be used by a physician for diagnosis of physiological difficulties. A test in addition, let us say, is devised in such a way that the half-dozen abilities which go to make up what we call the *ability to add* are tested separately. Thus we may have distinct tests for knowledge of the table, adding to a partial sum, bridging the tens, carrying from column to column, etc. By diagnostic tests of this sort the teacher is assisted in learning *why* a child cannot add, or subtract, or read, or write, and discovers exactly where he most needs help. Most of our instruction is of the hit-and-miss type, hitting the pupil's difficulty just occasionally and by happy accident. In medicine only the quack doctor gives medicine without finding the cause of the illness. In education, we nearly all do it. But by the help of diagnostic tests the teacher now has an opportunity to base her instruction on exact knowledge of each pupil's difficulties.

Standardized tests also show the *pupil* where he stands with reference to other pupils. He can compare his score either with that of other pupils in his room or with the average in America. The effect is excellent. He comes to feel that he is no longer working to meet the half-understood and half-accepted standard of the teacher, but is working to do as well as other pupils of his age or grade. This to him is definite and reasonable. He will spend hours in practicing with a football or baseball to be as proficient as other boys of his age. Toward fixed and intelligible standards in his studies his attitude is much the same. To be as good in a thing as the "other kids," especially if that thing is something which can be mastered by assiduous practice, is a motive that will arouse to strenuous exertion many a child who is now killing time.

Psychologists keep telling us that, given a pupil of a certain degree of native ability, the principal factor determining his rate of learning will be his resolution to learn,

his purpose to learn. Good teaching then requires, as its first factor, ability to arouse and maintain the purpose to learn. One of the best means of keeping fresh the purpose to learn is furnishing the pupil with definite objectives. The scores on standardized tests supply to the pupil's goal just this definiteness. When such scores are represented by a simple graph, say with one line showing the given pupil's attainment in these tests and another line the attainment of the average American child of this age or grade who has taken these tests, then the pupil has his strong and weak points set before him in a manner that is perfectly definite and objective.

The teacher will also find that by means of standardized tests she can very greatly increase the accuracy of her rating of the achievement of pupils. Our present methods of measuring the achievement of children in their studies are most regrettably defective. When teachers are asked to grade their papers a second time they sometimes miss their first mark by as much as fifteen or twenty per cent. A considerable number of teachers who tried this experiment in freshman classes at the University of Wisconsin found these extremes, and found that in general the amount of difference between the first and second marks averaged about five and a half per cent. Yet pupils glory in beating a rival by one or two per cent, and teachers debate with themselves whether to give a paper 88% or 89%—neither realizing that if the paper were graded again in a few days its mark would probably differ from this by some five per cent. When several teachers grade the same paper, the variations are still wider. A facsimile of a student's paper in an examination in geometry, graded by the principal teacher of mathematics in each of 110 accredited high schools in the Middle West, received marks ranging from 28% to 92%; and all the way from about 45% up to about 85% (except just at the passing mark) the marks were spread quite evenly.

The distribution of high and low marks within a class is equally erratic. Certain teachers give from three to ten times as many high marks as others in the same school. In a study of the office records for a considerable number of years in one institution, "A's" were actually found to be thirty-five times as common in one subject as in another, and "failures" to vary from an average of 33% in one subject to zero in another. Differences of this kind are now known to be the rule and not the exception.*

Are such marks likely to serve long in arousing pupils to effort? It is only too evident to pupils and students that marks do not depend upon achievement. Is it any wonder, then, that the relation between teacher and student, in the upper schools, is so commonly shot through with hypocrisy, and that it is so extremely difficult to establish normal social intercourse in the class-room? The students are students of the marking basis of their new teacher—of his or her whims and foibles and hobbies and pet aversions, and they "strive to please." Never yet has the discovery of truth proceeded successfully in any such atmosphere. And never will the desire for achievement appeal to those in the class most needing this motive until achievement is more fairly and more accurately measured.

Standardized tests do this. Standardized tests, because their scoring plan is standardized, are independent of the peculiarities of the person scoring them. All answers in the best of these tests are either right or wrong, and the score is the same whether the number of correct answers is counted up by one teacher or another, or by a clerk. In this way the pupil is credited with exactly the amount accomplished. When this amount is compared with the amount accomplished by pupils elsewhere, or by other pupils of the same ability and training in this school,

* On the inaccuracy of school marks, see Monroe, *Measuring the Results of Teaching*, Chapter 1; or Finkelstein, *Teachers' Marks;* or Rugg, *Teachers' Marks*, in *Educational Administration and Supervision*, Vol. 1.

marks or grades can be given accurately and dependably. Whether a pupil is ready for the next grade or the next course in the subject is then really known, not left to subjective estimate. And if tests of general intelligence are used in connection with tests of school achievement, pupils may be graded not always on a comparison of their achievement with that of other pupils of their age, but sometimes on a comparison of their achievement with that of other pupils of their degree of native ability. When marks are given according to the ratio of achievement to ability, dull pupils will no longer be subjected to proddings and ridicule in school and to floggings at home, all for failure to do work which they are really unable to do; nor will bright pupils need to be allowed to idle away their time in school and acquire habits of indolence which will handicap them throughout life: all children may be asked simply to keep their achievement up to that of the average child of that capacity.

By means of standardized tests, teachers as well as pupils may be rated more justly. A teacher's rating for efficiency ought to depend upon the progress of her pupils, due account being taken of the pupils' native ability. Given a class of a certain average ability, the teacher shows her skill by the difference in scores made on standardized tests by the pupils at the beginning and at the end of their stay with her. When lawyers and carpenters and engineers are judged on the basis of results secured, should a teacher continue to be judged on the basis of her "presence," or her voice, or her handwriting, or her use of methods favored by her superior? Is it not infinitely preferable that efficiency be proved by an exhibition of results secured? By standardized tests results are made tangible and measurable and are removed from the realm of debate. Measured results will make possible a genuine merit system of teacher promotion, with all the stimulation to effective lesson planning which that implies. It is

doubtful whether we have well-standardized tests in a sufficient variety of subjects to make this plan practicable at present. But when it comes it will certainly be the most effective single influence for the improvement of skill in teaching.

A teacher taking charge of a new room will find standardized tests useful in informing her of the educational status of her new pupils. By knowing this she can avoid re-teaching some things that the pupils already know, and teaching others for which they are not yet prepared. Exactly how much a class knows about a given topic as compared with the standard for that grade can be learned by using tests which have been given to thousands of other children in this grade, and which have scoring devices that make the scores independent of one's inevitably varying conceptions of scholarship.

The principal will find standardized tests valuable in correctly placing pupils who have been transferred from other schools. Very commonly such pupils are put back a grade for safety's sake. This injustice to the child can be avoided by using tests carrying with them a standard for each subject in each grade.

School officials also find advantage in using standardized-test scores as a medium for informing the public of the progress and the needs of the schools. Very frequently the public cannot understand what the schools are trying to do and the schools cannot tell them. Test scores furnish the common language, for anyone can understand what is meant by saying that our schools in Smithville are a year ahead of most schools of America in, say, arithmetic, and a year or two years behind others in music or French or manual training. The need of expansion of equipment for certain lines of work then becomes plain. School officials have also used standardized tests to refute unwarranted attacks on a certain school system by those who were politically interested in securing a change of

administration. And, of course, modern school "surveys" draw no conclusions about a school system without founding opinion on measured results.

Foreign countries regard the measuring movement as the most distinctive and significant feature of American education. Emissaries are sent to America to study it. Other countries are beginning to utilize the methods worked out here, and standardized tests are now in limited use in Western Europe, India, China, Hawaii, and Australia.

We ought to remind ourselves that the use of scientific method—the production of work that is precise, objective, impartial, and verifiable—has in its application brought all the comforts and conveniences of modern civilization. Scientific method has, for example, completely transformed agriculture in one century, in this time making greater changes in that industry than were effected in the preceding fifty centuries. Through its use of scientific method education seems likely to undergo the same transformation. Present school plans may in a short time seem as antiquated as the use of the sickle and flail in farming do now.

It behooves the teacher to become familiar with standardized educational measurements, both because of their probable future influence and because of the direct assistance they can render now in solving everyday classroom problems.

CHAPTER II

THE DIFFERENT KINDS OF TESTS

The principal kinds of standardized tests may be shown in outline form thus:

I. Intelligence or mental-alertness tests.
   1. Individual tests.
   2. Group tests.
      (a) Tests of abstract intelligence.
      (b) Tests of mechanical intelligence.
      (c) Tests of social intelligence.

II. Achievement tests.
   1. Research tests.
      (a) Primarily for comparison.
      (b) Primarily for diagnosis.
      (c) Primarily for prognosis.
   2. Practice tests.
   3. Teacher-made objective tests.

III. Intelligence-achievement synthetic scales.

IV. Miscellaneous.
   1. Scales measuring will-temperament.
   2. Scales measuring growth in religion.
   3. Scales measuring habits of good citizenship, etc.

I. **Intelligence tests.** The intelligence or mental-alertness tests are tests of native mental ability, or at least of such ability apart from the direct influence of schooling. They are presumed to show, in their common-

est form, the ability of a child to profit from his schooling. *Individual* intelligence tests are tests which much be given to one child at a time. The best-known example is the Binet-Simon test, which, in various revisions, is used in most large-city school systems for the discovery of "subnormal" or feebleminded children. *Group* intelligence tests are tests which can be given to a whole group at one time. The best-known of these is the Army Alpha test, devised during the war for the classification of the American recruits; but we now have group tests for almost all ages of children and for all degrees of intelligence in adults.

Tests of *abstract* intelligence are tests which measure one's mental ability by measuring one's ability to deal with abstract symbols, such as letters or words or numbers. The reality itself, the actual object, is not present—only the symbol representing it, such as its name or some letter or number to stand for it. For those not ready in dealing with abstractions, we are now beginning to have tests built from other sorts of material. Tests of *mechanical* intelligence are tests of one's ability to deal with machines. The test usually consists of a number of pieces of machinery taken apart, the requirement being to put them together again. The simplest bit of machinery may be nothing more than a nut and a bolt, the most intricate ones something as complex as the parts of a clock; and there are all grades of difficulty between.

Besides tests of abstract intelligence, or the ability to deal successfully with symbols, and tests of mechanical intelligence, or the ability to deal successfully with machines, there probably ought to be tests of social intelligence, or the ability to deal successfully with people. There are individuals who, it seems, learn but little from books, and who have no special aptitude for machinery, who can succeed as salesmen or business executives or in similar work whose primary demand is the ability to understand and

direct other people. These tests, however, are only in their first stages.

II. **Achievement tests.** The achievement tests measure, not the ability to learn, but the amount that has been learned. They are concerned with the pupil's proficiency in his school studies. Practically all of these are group tests. The *research* tests are those whose primary function is measurement, while the *practice* tests are those which measure results only incidentally and whose primary function is the improvement of results; that is, the practice tests are in their first intention teaching devices. Practice tests in arithmetic, for example, are skillfully graded outlines of "seat work," allowing each pupil to go forward at his own rate in mastering the four basic operations, and permitting him to prove his mastery of them by passing increasingly difficult tests. Of the research tests, some are designed primarily for the comparison of one school with another or with a standard of attainment, or for the comparison of the proficiency of pupils at one date with their proficiency at an earlier or a later date. Others are designed primarly for diagnosis, that is, for finding the reason that certain pupils do not learn. They test the pupil's ability in each part of the operation or topic separately, and attempt to discover by such analysis just where his difficulty lies. Others of the research tests are designed for prognosis, or forecasting the pupil's ability in a certain study.

The teacher-made objective tests—which are too new to have been adequately named—are statements about the material recently covered in any course, made up so that the answers desired can be indicated by underlining a word or by some other simple method which will make all the answers either right or wrong. One type of such a test shows four or more possible ways of completing each statement, and the pupil's task is to underline the one word which completes the statement correctly. For

example, "Chicago is in (Ohio, Indiana, Wisconsin, Illinois, Missouri)." Here the pupil's knowledge would be shown by underlining the word Illinois. Another type consists of a large number of statements so selected that about half of them are true and about half are false. The pupil shows his knowledge of the subject by writing the word *true* or the word *false* before each statement. These "home-made" objective tests are standardized in the sense that the scoring of them is done by a standardized plan which makes the result the same for all persons scoring the paper, but they are not standardized in the sense of having standard scores for comparison.

III. **Synthetic scales.** The intelligence-achievement synthetic scales are combinations of mental-alertness and school-attainment tests which allow one to measure at the same time the native ability and the school proficiency of a group of pupils—to get, at any rate, a general survey of the group. The Illinois Examination, one of the best-known of these, measures intelligence, silent-reading ability, and ability in the four fundamental operations in arithmetic. Such composite tests are frequently used as the basis for special promotions or for reclassification of pupils.

IV. **Miscellaneous.** Tests of will-temperament attempt to measure the so-called dynamic traits of personality—the endowment other than intelligence which makes for success. They are bringing useful results in certain schools training for business careers, where they assist in deciding the particular kind of business to which a given student may be best adapted. The other tests of this group are hardly far enough advanced to be considered reliable or really standardized.

The dependable and important tests at present are the tests of intelligence or mental alertness and the tests of school achievement.

CHAPTER III

TESTS OF ABILITY TO LEARN

The Meaning of General Intelligence

Intelligence has been variously defined as the ability to learn, the ability to carry on abstract thinking, the ability to adapt oneself to new situations, the ability to use one's mental powers in a productive way. The latter definition is intended to include something more than brightness or mental alertness, for, as its author points out, one may be bright and alert and yet not manage his affairs well. Such a person may lack mental *balance,* or he may lack the power to resist suggestion, or the power to see any but commonplace relations among experiences, or the power to see which, among a great mass of facts presumed to bear on a question, is the single significant fact. Admitting that such are the characteristics of the superior individual and that they are of the first importance in practical and in scientific work, we may well question, however, whether they are the traits with which we are most concerned in school. For, much as we may regret to say so, the kind of intelligence most required for success in school work is brightness or alertness. Therefore, so long at any rate as we are primarily concerned with measuring intelligence for the purpose of predicting school success, it is unnecessary to define intelligence so broadly as to include these traits, admirable as they are.

Defining intelligence as the ability to adapt oneself to a new situation again emphasizes a trait which is not of outstanding importance in school work. In most schools

the major part of the adapting is done for the pupil by the teacher. Seldom is the child asked to go up against a really novel situation, for the introduction to the new type of problem is nearly always given him as a part of the instruction. Although the ability to deal with and adapt oneself to a new situation may be ever so desirable, it is not precisely the ability most needed for school success.

Defining intelligence as the ability to carry on abstract thinking comes nearer to the type of intelligence now required in school. But this somewhat overemphasizes the element of thinking. In perfecting skills, such as writing or reading, or adding, not a great deal of thinking is required. Yet the child who, if he tries, can rapidly master these things would usually be called more intelligent than the child who cannot. Therefore to limit intelligence strictly to the ability to carry on thinking or reasoning would seem to make the definition a little too narrow.

"Ability to learn" is probably the most satisfactory brief definition of intelligence for the teacher. It is ability to learn which must be taken into account in almost any kind of school experiment. It is exactly the ability upon which the work of the school depends. And it includes not only thinking as dealing with novel tasks, but the learning of school tasks of all sorts. Whether there is such a thing as *general* intelligence, and whether this question is an important one at present, will be discussed on page 23.

Composition of the Tests

Intelligence tests were first successfully worked out when, after the persons in charge of the schools of Paris had established special schools for subnormal children without providing a method for selecting such children, the psychologist Alfred Binet attempted to perfect devices by which subnormal children could with certainty be discovered. For it appeared to M. Binet a very serious

thing to designate a child as feebleminded, and the cause of a great injustice if a mistake were to be made. Binet, with the assistance of a physician, Simon, worked on a new principle. Previously, many "mental tests" had been devised in psychological work, but they were usually tests of the simpler mental processes. They measured intelligence, if at all, only by measuring something supposed to be correlated with it. Binet began by attempting to measure intelligence directly. He devised tests the solution of which required thinking and judgment. Furthermore, he used tests not singly, as heretofore, but in groups. Binet's tests really worked, in the sense that when they were given to children whose intelligence was already known by long association, they placed the children in correct order. It could then be assumed that they would rate children unknown to the examiner correctly in comparison with children of a known degree of intelligence. Measuring intelligence directly and by groups of tests was thus proved superior to measuring it indirectly and by single tests.

Binet's other great contribution to mental measurement was to introduce the idea of age levels—the idea that at each age the average child is able to solve a certain number or a certain type of problems, and that the intelligence of any given child can therefore be expressed in terms of the "mental age" which he is thus shown to have reached. A group of tests could then be standardized as the tests which should be passed by a child of a given age if his intelligence was average; and his acceleration or retardation in mental growth could be expressed in terms of years.

Binet published his first set of tests in 1905 and revised them in 1908 and again in 1911, and further revisions were made by Goddard, Terman, and others in America. Terman's Stanford Revision of the Binet Scale is now the most widely used intelligence scale for careful individual measurement, and is commonly regarded as a

remarkably accurate psychological instrument. It consists of ninety tests, with six or eight at each age level from three years to sixteen years. For the youngest children, the tests are of such simple things as pointing to the nose, eyes, etc.; naming familiar objects, such as a knife, key, penny; telling whether a boy or girl, and so on. Older children are asked to define such abstract words as *pity, revenge, charity;* to fill out words in dissected sentences; to discover the meaning of fables, and so on. The standards for this scale are the results of very careful work with thousands of children. Its use should, however, be left to the expert, since accurate and reliable results can be secured only after the examiner has undergone considerable training. So much for individual tests.

The *group* tests of intelligence are one of the outcomes of the war. Since it was desirable to test the intelligence of all recruits entering the army, and since this was obviously impossible if the men were taken one at a time, as the Binet method required, it was necessary to devise a test which could be administered to large numbers simultaneously. Such a test was perfected by American psychologists during 1917 and 1918, was given to one million seven hundred thousand soldiers, and after the war was given to a very large number of high school and college students. Adaptations to younger children were then worked out, so that today we have group tests for every age from six to sixteen—the age at which mental maturity is presumed to have been reached.

The group intelligence tests contain such problems as the following: The pupil may be given a paper covered with letters and geometrical figures and be asked to draw certain lines from one to the other, or to underline certain of them. This is a test of ability to carry out directions. The directions are given rather rapidly and must be executed in a limited amount of time. A second test may consist simply of arithmetical problems, ranging in diffi-

culty from such simple exercises as "How many are 30 and 7 men?" to "If a man runs a hundred yards in ten seconds, how many *feet* does he run in a fifth of a second?" A third test may be a test of common sense, such as placing a cross before the best reason of the following three: "Why do we use stoves? Because ( ) they look well, ( ) they keep us warm, ( ) they are black." Another test may ask the student to indicate whether a certain pair of words have the same or opposite meanings; or to show the relationship between words by some such scheme as underlining the word in italics which bears the same relation to the third word that the second does to the first, in the following series: "Gun—shoots; knife—*runs, cuts, hat, bird.*" Or the test may ask the person examined to add the next two numbers to a series such as: "3, 6, 9, 12, 15, 18, —, —." It may ask him to point out which four of a series of six pictures are alike in some way; e. g., they all refer to summer, or all to a certain kind of act. It may simply ask for items of general information, such as would be shown by underlining one of the words in italics in this sentence: "The tuna is a kind of *fish, bird, reptile, insect.*" Each of these tests usually begins with very easy questions and goes on gradually to very difficult ones. The time limit prevents anyone but a genius from finishing the test, and the scores are computed in terms of the number of exercises that have been completed correctly in the given time.

The underlying principle of such testing is that we can get a fairly good measurement of a person's *general* intelligence by taking, as it were, samplings here and there of different kinds of ability. Whether this method actually measures intelligence can be determined only by selecting a certain number of persons to test and checking up the results against some slower and presumably more accurate measurement of intelligence, such as that secured through long acquaintance with the selected individuals. To what

extent the tests are thus vindicated is discussed below under "validity of the tests." It seems obvious that if a thirty-minute intelligence test puts a group of men in the same order as that in which they would be placed by persons well acquainted with them, then the officer can know at once the kind of men he has to drill and the teacher can know on the first day of a semester—not after several weeks—the amount of ability of each child whom she is to instruct that semester.

Another principle on which such tests are built is that if intelligence means ability to learn, it can be measured not only by having the tested person learn something new during the test, but also by measuring the amount he has learned in the past. It is on this basis that the intelligence scales justify the use of tests of general information, and, to a certain extent, the tests involving arithmetical problems. The items in such tests must be selected from sources of information or from types of training which are common to all the persons tested. They must not, of course, be drawn from special fields of learning or from special kinds of environment. In the Army tests they were apparently taken from items learned in the first four or five grades of the schools and from the reading matter and advertisements of newspapers. Similar considerations have controlled the making of intelligence tests for use in schools.

Validity of Intelligence Tests

If intelligence is taken as ability to learn, then the fact that the existing tests usually presuppose a certain educational background and a certain ability to deal with abstractions is no drawback as far as ordinary school uses of the tests are concerned. The tests may indeed be to a considerable extent linguistic, but linguistic ability is probably the most important single factor in success in the present-day schools. We are ordinarily interested in learn-

ing whether given pupils have or have not the ability to master the present school tasks. Experimental evidence shows that the intelligence tests now available are able to measure this ability very well indeed.

The fact that many men of proven ability failed in their school work—Oliver Goldsmith, Lord Byron, Charles Darwin, etc.—might imply either that the ability of such men is of a highly specialized type, or that these men found nothing in the schools of their day which appealed to them as worth while. The latter theory will explain such cases, but will not so readily explain the cases of boys, known to all of us, who do try to master their school tasks, without much result, and who later become successful mechanics and business executives. These persons seem to be endowed with special ability to deal with machines or with people, even though they cannot deal with anything abstract. Our ordinary intelligence tests do not measure such abilities. A special test has recently been designed for measuring mechanical aptitude,[1] but the test for measuring social aptitude is still in the future. The existence of what appear to be these three types of intelligence—the abstract, the mechanical, and the social—may necessitate the reorganization of the schools to provide more adequately for the latter two, but it does not invalidate the work of the present intelligence tests in measuring the ability to master the present school curriculum.

Another point a little difficult to keep clearly in mind is this: Saying that intelligence tests measure a pupil's ability to master school work is not saying that they measure the probability that he will master it. Purpose and effort count heavily, of course, and general intelligence tests should not be criticized for their failure to measure factors such as these, which they were not designed to measure. Intelligence tests will not tell us which pupils will succeed, but only which pupils will suc-

[1] *Stenquist Mechanical Aptitude Test.*

ceed if they try. The latter is obviously a very valuable kind of information, since by means of it teachers and parents can with confidence put on pressure to make sure that indolent pupils do try, and can likewise lessen the pressure on pupils shown by the tests to be doing all they are able to do.

That intelligence tests really do measure intelligence as well in one hour as it can be estimated after several months' acquaintance is proved by methods such as giving the tests to pupils whom the teacher knows well, and comparing the order in which the pupils are placed by the tests with the order in which they are placed by the teacher. For most of the children in the class the agreement will be remarkably close. If one then goes a step farther and studies very carefully the pupils about whom the teacher and the test do not agree, he can discover which of the two is the more dependable. Such experiments have been carried on in a number of cities and sometimes with large numbers of pupils. For example, in the junior high school of the University of Oregon[2] one hundred and twenty-five pupils were given three different intelligence tests and at the same time were ranked as to their intelligence by six of their teachers independently. The teachers' estimate of intelligence was made out in this case with a great deal of care. A study of the cases regarding which the teachers and the tests did not agree then showed that the teachers tended to overestimate over-age pupils and pupils who were talkative and vivacious, and to underestimate the younger and physically-undeveloped pupils or pupils who were shy and retiring, or else that they made no distinction between a pupil's intelligence and his proficiency in his studies. Studies made earlier by Binet and by Terman had had the same outcome.[3] Such experiments would seem to show that

[2] Ruch, *Study of Mental, Pedagogical, and Physical Development,* University of Oregon Publications, No. 7.

[3] Terman: *Measurement of Intelligence,* Chapter 2.

the scores on a brief intelligence test are often even more dependable than the pooled opinion of several teachers well acquainted with the children.

Intelligence tests can also be checked, though less satisfactorily, by comparing their results with school marks. When this was done, for example, in the High School of Leavenworth, Kansas,[4] the agreement was found to be especially close in such abstract subjects as Latin and algebra. Forty-five cases in which test scores and school marks did not agree were then investigated individually, and in all except three cases the low marks of these pupils were found to result from poor health, indolence, irregular attendance, or some other factor besides low intelligence, and the high marks from such influences as excellence of attitude in class or exceptional effort. In other words, the tests were shown as before to be the more accurate measurements. Similar experiments brought similar results among the students of Brown University[5] and Smith College.

Intelligence tests were checked in the Army by comparing their rankings of the men of a company with the rankings given by an officer who had known the men for several months. Thus, in one group of over seven hundred men, whose officers were asked to rank them as to "practical soldier value," there was substantial agreement between officers' rating and test-rating in 88% of the cases. Considering the number of factors which influence practical soldier-value besides intelligence, this seems remarkably close. About the same amount of agreement was discovered independently in several other camps, in experiments each of which involved several hundred men.

Intelligence tests can also be checked by seeing whether the highest scores are made by persons in occupations

[4] Bright: *Intelligence Examination of High School Freshmen.* Journal of Educational Research, June, 1921.

[5] Colvin in *Educational Review*, June, 1920; *School and Society*, July 5, 1919, and July 29, 1922.

which are commonly supposed to require the highest intelligence. When army-test scores were analyzed according to the occupations of the men taking the test, the order of occupations by size of score was this: Professions, clerical occupations, trades, partially skilled labor, unskilled labor. This is the order of intelligence in which any schoolman would put the occupations if intelligence is taken in the sense of the school ability of the boys who later enter these occupations. Similar experiments at Leland Stanford University, using from thirty to two hundred representatives of each occupation, placed the groups in the following order: College students (future followers of the professions), business men, express employes, motormen and conductors, firemen and policemen, salesgirls. More work of this type, to supplement the army results, is very much needed.

The best way, however, to check the validity of intelligence tests for school use is to try them out in the schools. Their commonest use is for purposes of classification of pupils into groups of uniform ability. The plan is in effect, for example, in the Harrison Technical High School, Chicago; the University of Minnesota High School, and in the high schools of Montclair, New Jersey; Long Beach, California, and Oakland, California. From each of these schools the experiment is reported a success. Some of the schools have stated that they could not be induced to return to the old haphazard method of classifying pupils. As to the reliability of scores, the group tests are thus proved to give results which at the least are sufficiently accurate to effect a very great advance over existing methods of classification.

The score on a single group test of intelligence should not be depended upon, however, where a decision is to be made about a single individual which will greatly influence his future. To decide, for example, whether a certain pupil is to be considered feeble-minded and sent

to a special room or a special school, an individual test should be used. And for making less momentous but still important decisions about a single individual, the results of two or more group tests should be combined. The group-test score for each individual should always be considered an approximation. In measuring the intelligence of a *collection* of individuals, such as those of a school or a room as a whole, the group tests are much more thoroughly dependable, for with a large number of measurements the small errors in each measurement tend to balance and cancel each other. The average or median scores resulting from the use of a group intelligence scale may therefore be accepted as accurate, while the score of each individual should be considered a rough measure subject to some correction from other intelligence tests which may be given later.

"Do not speed tests, or tests with a time limit, penalize unjustly the person who thinks slowly and accurately?" one is frequently asked. "If more time were given, would not the slow thinker often prove the most intelligent of all?" The evidence we have so far secured seems to prove the opposite. When the work of persons who have answered only a few questions of the test is compared with the work of persons who have covered much more ground, the former are discovered to have made more mistakes than the latter. This is true in spite of the fact that the slow worker may have finished so little of the work that he had perhaps only half as many chances to make a mistake. As to extending the time, this was tried out pretty thoroughly in the Army. In one camp 123 men, in another 387, and in another 510 men were given the intelligence test, first with the usual time limits and then with the time doubled. The ranks given the men by the two methods of testing them were almost exactly

the same.[6] In fact, this was one of the closest confirmations ever found in work with intelligence tests.

The Principal Uses of Intelligence Tests

The bright children of each grade may be placed together in one room, as explained before, the mediocre children in another, and the dull children in another. The bright children may then either be helped to cover the required work more rapidly and to finish their course more quickly, or they may be put through the course in the same number of years as others, but be given an enriched curriculum each year. Similarly, the dull pupils may be allowed either to cover the regular course in a longer time or to finish in the regular time with minimum essentials only.

The latter plan is in effect, as one instance, in the elementary schools of Detroit. There the children are divided upon first entering school into a bright section composed of the upper 20%, a medium section of the middle 60%, and a dull section of the lower 20%, who are given respectively the enriched course of study, the regular course, and the simplified course. Although the test classification is considered tentative, and shifting pupils from one group to another is permitted, very little shifting has been necessary.[7]

Uniform classification permits adaptation of the teaching method to the ability of the pupils. For example, more time can be given to drill and to review in the class of dull pupils, and this cuts down the number of failures. It gives some of the dull pupils their first chance to develop as leaders in the classroom. It eliminates, according to many teachers who have tried it, a large number of disciplinary difficulties—for the dull pupil can understand

[6] *Psychological Examining in the United States,* Part II, Chapter 9. Official Report to the Surgeon General.

[7] For further details of the working of the plan, see the *Twenty-first Yearbook of the National Society for the Study of Education.*

what is going on, and the bright pupil is not obliged to kill time. It gives some of the bright pupils their first taste of real competition. It thus prevents the formation of injurious and tenacious habits of indolence, and makes it possible to develop to their full capacity those on whom the country must depend for its thinkers in every line of endeavor.

Intelligence tests can be used, even where classification is impossible, as a basis of school marks. Many schools are beginning to give marks and promote pupils, not on the basis of what the pupil accomplishes as compared with other pupils, but on the basis of what he accomplishes as compared with his ability. Ability is measured, of course, by intelligence tests, and achievement by achievement or subject-matter tests. This plan is found to bring real effort from almost everyone and to overstrain no one.

Intelligence tests may also be used, in connection with achievement tests, in deciding upon the efficiency of a given piece of teaching. The intelligence tests tell what the pupils are capable of doing, the achievement tests tell what they have actually done. A good deal of light can be thrown in this way upon the value of different methods or devices in teaching.

In high schools and colleges, intelligence tests may be used in determining which students are to be permitted to carry extra courses. The older method, which makes this decision depend upon average scholarship in the previous semester, often induces the more ambitious but less hardy students—especially adolescent girls—to overwork, and bring permanent injury to their health. Intelligence measurements permit extra work to be carried only by those who are bright enough to do it without injustice to themselves. Besides determining the amount of work, the tests may also determine the kind of work a student should attempt. That students below a certain standing on the intelligence scale will almost certainly fail in the more

abstract subjects, such as algebra and Latin, we know because it has been observed that practically all students below this degree of intelligence do fail. Unless we believe very thoroughly in the superiority of the abstract type of intelligence, this is not necessarily a reflection on the student, and in advising him away from those studies we need not present it as such.

Both in the junior and in the senior high schools, the intelligence tests are useful as a basis for advice in the choice of a vocation. Although we cannot tell a boy the occupation in which he is pretty sure to succeed—for this depends largely upon temperamental and emotional traits—we can tell him many of the occupations in which he is practically certain to fail; for each occupation requires a certain minimum degree of intelligence. The professions, for example, cannot be entered without graduation from high school and from college, and the degree of intelligence for graduation from college, as well as the degree of intelligence for graduation from high school, is pretty well known in terms of scores on intelligence tests taken very much earlier. We know in a rough way the degree of intelligence characteristic of each of the occupations reported by the two million men examined in the United States Army. The occupations whose members made the lowest average scores were those of: laborer, general miner, teamster, barber; those in the next group were: horseshoer, bricklayer, cook, baker, painter, general blacksmith, general carpenter, butcher, general machinist, hand riveter, telephone and telegraph linesman, general pipefitter, plumber, tool and gauge maker, gunsmith, general mechanic, general auto repairman, auto engine mechanic, auto assembler, ship carpenter, telephone operator; those in the next higher group were: concrete construction foreman, stock-keeper, photographer, telegrapher, railroad clerk, filing clerk, general clerk, army nurse, bookkeeper. The next higher group included:

dental officer, mechanical draftsman, accountant, civil engineer, medical officer. The highest group included army chaplains and engineer officers. It is true that vocations are now very commonly taken up for reasons other than one's special fitness for them, and thus that many of the men in these occupations do not really belong in them. But a considerable number of bad choices are remedied by shifting about from one occupation to another, and the above list may be taken as furnishing at least a rough indication of the amount of (abstract) intelligence characteristic of the large groups of occupations. On this basis some advice can be given, particularly as to the vocations that ought not be attempted. It is important that school people give up their present inclination to advise all pupils to be ambitious enough to try to climb into the more intellectual occupations. This advice is injurious both to the boy who may be thus led into a failure in life and to the clients or patients whom he may but half competently serve. If occupations are considered respectable and admirable in the degree to which they meet real and wholesome needs among the people they serve, and not, as at present, in the degree to which they involve the use of the brain rather than the hand, then the whole basis of vocational advice will be changed. Intelligence tests may be made to assist very greatly in placing the pupil where he will render his best service.

In colleges and universities intelligence tests are coming to be used as at least a partial basis for determining admission. They are so applied, for example, at Columbia University, the Carnegie Institute of Technology, and the University of Michigan. They are also used in advising a student in choice of studies, as at the School of Commerce and Administration of the University of Chicago. Sometimes they are used in deciding whether a student is to be asked to withdraw from the university, for it is obvious that if his low grades in first-semester work are

coupled with very low intelligence scores it is useless to ask him to stay longer, while if they appear in connection with very good intelligence scores, the cause of the low grades may be something remediable, such as inadequate preparation or poor study habits. In certain colleges intelligence scores are used as a basis for grades, and the brighter students are compelled to keep their attainment up to their ability. In at least one college, intelligence tests are likely to serve as the ground for the award of a very large number of scholarships. The general effect of the measurement of intelligence in colleges seems to be the improvement of the morale of the student body, for the students come to feel that the faculty is not so much engaged in exacting work from them for a diploma as in guiding and helping them to make the most of their abilities.

Outside of schools, tests of intelligence are coming to be used in connection with such industrial and social questions as immigration, vocational placement, and the treatment of criminals. In deciding whether an immigrant is to be admitted to this country, it seems more sensible to test his intelligence than to test his degree of literacy. For illiteracy may be an accidental handicap and certainly is a remediable one, while low intelligence is not only a permanent characteristic of the individual, but will probably be a characteristic of a large number of his offspring for all time. Those of low-grade intelligence, not the illiterates, are really the least desirable citizens.

For vocational placement the tests of intelligence already are widely used, and by the employment departments of many large companies they are considered standard equipment. They are used both to determine whether an applicant shall be accepted, and, if accepted, to determine for which kind of work he shall be trained. To place a man in a job in which he will be contented, and thus to reduce the labor turnover, is rapidly coming to

be thought of as good economy. Experiments in industrial establishments have shown that there is a very close relationship between one's liking for his work and his intellectual adjustment to it. Neither able men in routine work nor dull men in highly organized work report themselves as liking their jobs. On the other hand, both able men in difficult work and stupid men in simple mechanical work report that their work is to their liking. Work adapted to the ability of the individual, other things equal, means a longer stay at the job and a lessened cost in breaking in new men.

From the point of view of the employee, as well as from that of the employer, intelligence tests for job placement are of great value. There is no more real contribution to one's happiness than to spend the eight or ten hours of one's working day in activity which one enjoys. The working hours make up the largest continuous section of one's waking life. School people must come to see that to assist a pupil to choose a congenial life work is the greatest single benefit they can confer on him. Intelligence scores, although by no means the only criterion in selecting a vocation, are yet one of the most important of the criteria.

Criminology makes increasing use of intelligence tests in deciding responsibility for infractions of the law. It is to be hoped that it will soon be possible to do away with the present common method of dealing with cases of doubtful responsibility, that is, the giving of a compromise sentence. If the criminal is feeble-minded he should indeed be placed where he is not likely to commit another crime, but he should not be punished. If he is of normal intelligence and capable of understanding the effects of his acts, then he should be punished so severely as to deter others from crime. The compromise sentence defeats both these objectives.

Such are some of the uses of the intelligence tests.

Their abuses, unfortunately, would furnish an almost equally long list. The testing movement has grown so rapidly that its greatest danger now is in its over-enthusiastic friends. An intelligence test is not a panacea for all school ills; neither is it fool-proof. It is necessary that users of intelligence tests know something of what is being measured, and something of the interpretations that may or may not be put upon the resulting scores, or these scores are sure to be misapplied. It is well to remember how rough a single measurement of an individual is. And it is well, in making applications of the results, to remember that the kind of intelligence test in most common use probably is not a general intelligence test at all, but a test of one particular type of intelligence, the academic type; and that, therefore, however well the tests acquit themselves in applications which lie within the school, they must at any rate be used very much more cautiously in applications which reach outside the school.

The future influence of perfected scales for measuring intelligence is something we cannot yet imagine. It is sometimes said that our manner of life was completely revolutionized in the nineteenth century by the utilization of power machinery, and that it will again be revolutionized in the twentieth century by the utilization of air transportation. But it is possible that intelligence tests may bring the race more real progress than either of these, because they will serve to place in strategic positions the men most capable of using the opportunities for progress which lie there. They will put able men where ability will count. Advancement can thus be tremendously accelerated.

Selection of an Intelligence Test

There are a few very simple criteria to keep in mind in choosing an intelligence test for a given experiment, and others not so simple. Of course, one should consider the pupil-age for which the test was designed, in order

that he may not perhaps secure for upper grades a test designed for primary children or for college students. He should also consider the time taken to administer the test, with reference to the length of his class period, or to the possibility of extending the period. He should consider the length of time needed for scoring the papers. Many tests cut this down by special devices, such as transparent answer sheets to be laid over the pupil's answer list. He should, of course, consider money cost, though the amount of difference between the various tests in this matter is not ordinarily great enough to make this a very important item.

Less simple considerations are the extent to which the given test has been used and the degree to which its results check up with other tests or with a repetition of the same test. The extent to which a test has been used shows something of the number of cases upon which its mental-age norms are based, and shows in a very rough way how effective it has proved under actual schoolroom conditions. But this is, of course, an unfair criterion for the newer tests, and the newer ones are often best because their authors have had an opportunity to profit from pioneer mistakes. Furthermore, extent of use may depend more directly upon vigor of advertising by the retailing company than upon merit. The degree to which one intelligence test agrees with others, when four or five are given to the same children, and the degree to which a test agrees with a repetition of itself, are excellent checks. These can be found out by reading the author's announcement of the construction of his scale and the tests to which he has himself put it, and by reading the reports of persons who have subsequently procured and used the scale. These accounts will be found in the educational and psychological journals, particularly in the *Journal of Educational Psychology,* the *Journal of Applied Psychol-*

ogy, the *Journal of Educational Research,* the *Elementary School Journal,* the *School Review* and *School and Society.*

Another important consideration is the purpose for which the test is being given. If the purpose is the prediction of scholarship, or more strictly of ability to master the present curriculum in case effort is made, then the relationship already established between school marks and scores on the given test is important. That is, if we wish to find whether this test will measure the pupil's ability to learn what the schools now teach, it is important to see whether the ability it does measure has already been found by an examination of grades to be closely related to scholarship. Checks of this kind are desirable, for example, where the tests are to be used for college or high school entrance, for special promotions in the elementary school, or for classification of pupils into groups of uniform ability. On the other hand, if the purpose of the experiment is vocational advice, then the relationship between test scores and scholarship is of less importance. If the purpose is to test for commitment to a special school or institution, the individual rather than the group tests should be selected.

A list of all the standardized tests in print, both intelligence and achievement tests, is now available under the title, "Bibliography of Tests for Use in Schools," which can be procured for ten cents from the World Book Company. This booklet gives the name of the publisher of each test, and in some cases the reference to the journal in which the test was announced and described by its author. The forty-six intelligence tests published before 1922 are listed in the Twenty-first Yearbook of the National Society for the Study of Education (published by the Public School Publishing Company). This list gives the title and author of each test or scale, the number and nature of the different tests included, the age of pupil to which the scale is adapted, the number of minutes

needed for the test, the publisher and price, and the journal in which the test is described. Information about the available tests can be secured by writing for the price lists and descriptive literature issued by the publishing houses especially interested in this field, some of which are:

The Plymouth Press, Chicago, Ill.

The World Book Company, Chicago, Ill., or Yonkers-on-Hudson, N. Y.

The Public School Publishing Company, Bloomington, Ill.

The C. H. Stoelting Company, Chicago, Ill.

The Bureau of Publication, Teachers' College, Columbia University, New York.

Some of the best-known intelligence tests are the following:

For the kindergarten and the primary grades: The Detroit Kindergarten Test; Detroit First-grade Intelligence Test; Haggerty Intelligence Examination, Delta I; Holley Picture Completion Test; Kingsbury Primary Group Intelligence Scale; Otis Group Intelligence Scale, Primary Examination; Pressey Mental Survey Tests, Primer Scale.

For the intermediate and upper grades: Haggerty Intelligence Examination, Delta II; Illinois General Intelligence Scale; National Intelligence Tests; Otis Group Intelligence Scale, Advanced Examination; Pintner Mental Survey Tests; Pintner Non-Language Mental Tests; Pressey Cross-Out Tests; Whipple Group Test for Grammar Grades.

For the high school: Army Alpha Intelligence Tests; Miller Mental Ability Tests; Otis Group Intelligence Scale, Advanced Examination; Terman Group Tests of Mental Ability.

For colleges and universities: Brown University Psychological Examination; Roback Mentality Tests for Superior Adults; Rogers Group Tests of Intelligence; Thorn-

dike Intelligence Examination for High School Graduates; Thurstone Psychological Examination for College Freshmen and High School Seniors.

For all grades: The Binet-Simon Scale, and its revisions by Goddard, Yerkes, Kuhlman, and Terman; the Myers Mental Measure; the Trabue Mentimeter.

Giving the Tests

The person giving the test should remember that reliable results cannot be secured unless the printed directions are followed exactly as they are stated. Directions for giving the test are sent with the test blanks. A small change in the manner of giving the test may affect the scores so as to render them useless. If several rooms are to be tested, and the results compared, it is better to have all the tests given by one person, so as to have the conditions more nearly uniform.

The papers are next scored by means of the printed answer lists. The score may be written on the cover of each pupil's paper. The "point score," or sum of the points made on the various tests, should then be converted into mental ages. A pupil's mental age is the same as the chronological age of the average pupil who makes his score. An example will make this clear. If a child should make seventy points on a certain intelligence scale, and it should be found that seventy points is the score made by the average nine-year-old child, then his mental age is nine years. This means that he is as mature mentally as the average child of nine years. The mental age corresponding to each point score will be found written opposite the point score in tables furnished by the maker of the test.

The teacher may next wish to find the intelligence quotients. As the mental age is the measure of mental maturity or ability, the intelligence quotient is the measure of brightness, or ability in relation to age. The intelli-

gence quotient is found by dividing the mental age by the chronological age. For example, if a boy has a mental age of 12 and a life age of 10, his intelligence quotient (or I. Q.) is 1.20, or, if we drop the decimal point, 120. The I. Q. of the average child is, then, 100; that of duller children is below 100, and that of brighter children is above 100. The exact meaning of any given I. Q.—the comparative amount of brightness or dullness implied—is shown in tabular form for each test. The intelligence quotient can be seen to have a meaning quite different from that of the mental age if we compare a child whose mental age is 12 and chronological age 10 with another whose mental age is 6 and chronological age 5. Both will have I. Q.'s of 120 and both will therefore be equally bright, but the older child, because of his maturity, will be able to solve many kinds of problems and work his way out of many difficulties, in the face of which the younger child would be helpless. Their brightness is the same, but their mental ability is very different.

Many of the current uses of mental tests have been mentioned above, and some of the commoner abuses pointed out. Some definite use of the results ought to be in mind before the test is selected and purchased. Thousands of tests are given each year which lead to no change whatever in any part of the work of the school. Such a waste of the time of teachers and pupils can be avoided by careful study of the meaning and implications of the measurement of intelligence.

Caution has already been given about overconfidence in results when one is dealing with single pupils. The lowest scores, in particular, should be considered tentative. They may be the result of fright or temporary indisposition as well as the result of stupidity. It is well, therefore, to check these cases by giving a second form of the test to such pupils. For this purpose a small number of the second form may be ordered when the other test

blanks are sent for. About ten per cent should be subtracted from the second score in order to allow for the effects of the previous acquaintance with this type of test. A duplicate form of a test, it may be explained, is another edition of the test, made up of exercises similar to those in the first form, but not identical with them.

The teacher who wishes to become thoroughly acquainted with the literature on intelligence testing should subscribe for one or more of the journals mentioned on pages 35 and 36, and should read some of the following books:

The Twenty-first Yearbook of the National Society for the Study of Education. (The Public School Publishing Co.) Part I outlines the nature, history, and principles of intelligence testing, and Part II describes administrative uses of intelligence tests in various cities and for various ages of pupils.

Terman: *The Measurement of Intelligence,* and *The Intelligence of School Children.* (Houghton Mifflin Co.) Scholarly and readable accounts of the revision and application of the Binet scale.

Ballard: *Mental Tests.* (Hodder and Stoughton, London.) A very interesting outline of the subject by an Englishman. Includes description of new tests for discovering super-normal children.

Book: *The Intelligence of High School Seniors.* (The Macmillan Company,) Description of the measurement of the intelligence of six thousand seniors in Indiana, and the relation of the scores to scholarship, vocational preference, college choice, economic status, sex differences, etc.

Trabue and Stockbridge: *Measure Your Mind.* (Doubleday Page & Co.) A well-written general account of the significance of psychological measurements, and an outline of the "mentimeter" tests.

Yoakum and Yerkes: *Army Mental Tests.* (Henry

Holt & Co.) A very interesting portrayal of the psychological work in the United States army.

Terman *et al*: *Intelligence Tests and School Reorganization.* (World Book Co.) A brief general discussion followed by a description of the uses now made of intelligence tests in certain selected cities, large and small.

Goddard: *Human Efficiency and Levels of Intelligence.* (Princeton University Press.) Informal lectures, printed within about a hundred pages, discussing mental measurement in its social implications.

CHAPTER IV

TESTS OF AMOUNT LEARNED

General Statement

The development of achievement tests runs closely parallel to the development of intelligence tests. The achievement or educational tests reached a usable stage somewhat later than the individual tests of intelligence, but earlier than the group tests.

It seems that the first definite scale for measuring the excellence of school work was made by a certain Reverend George Fisher, and was described as being in regular use in the Greenwich Hospital School in England, in 1864. The Reverend Fisher constructed what he called a scale-book—a set of samples of pupil work in handwriting, grammar, spelling, drawing, etc.—which was kept on file to show the degree of excellence expected in each division of the school. No one saw the significance of the device at the time, and it was not copied elsewhere. In America the beginnings of standardized measurements of school products were made by Dr. J. M. Rice in 1895, when he put together a list of words which he applied in various schools as a spelling test. His surprising discovery was that children in schools which devoted very little time to spelling were able to spell as well as children in schools which spent many hours in drill on spelling. These results he published in the *Forum,* then a widely-read magazine, where as a news feature story they attracted for a short time considerable notice. He also presented his results at the annual meeting of the National Education Association, but there found the coldest of receptions. For the peda-

gogues were agreed that his scores showed nothing, since he was under a misapprehension as to the purpose of education. Dr. Rice, not being a schoolman, had supposed that the purpose of teaching spelling was to give the children the ability to spell. The schoolmen informed him that the purpose of teaching spelling was to train the pupils' minds!

After this first attempt at the scientific measurement of classroom results had been wrecked in the 90's against the dominating concept of formal discipline, nothing further of the kind was reported until 1908. In that year Stone published his scale of reasoning problems in arithmetic. The following year Thorndike published his handwriting scale, a graded series of samples of penmanship, by means of which a pupil's ability was to be determined by comparison of his writing with these samples. This third recurrence of standardized measurements in education was destined to live. In 1914 the movement had come sufficiently into favor to be endorsed by the National Education Association—no longer so completely under the influence of the theory of formal discipline—and since that time its growth has been extraordinary. At present it is receiving more attention and, better, more careful painstaking endeavor, than any other movement in education. Two leading journals are devoted entirely to quantitative educational studies, and three or four others are giving more than half their space to such studies. More and more of our graduate students are specializing in this field. As these highly-trained men and women procure for us more and more *facts* about pupils and teachers and schools, we may expect educational discussions to move more and more away from their past boresome character of moralistic injunction and unguarded speculation resting on the flimsiest of factual bases—the kind of speculation which leaves one with the impression that an equally able man could make out an equally good case for the other side of the ques-

tion—and to acquire gradually a set of doctrines which will stand permanently and which can be confidently accepted and used, because they are based on information that is precise, objective, impartial, and verifiable, or in other words, scientific.

It is sometimes believed that the use of standardized tests will tend to overemphasize the mechanical aspects of school work. This belief is natural, since the tests which were first developed, and consequently are now most widely used, are tests of the simple and definite results of "drill" work. As these are the outcomes easiest to measure, and so most commonly measured, teachers often may receive the impression that if they wish to appear efficient they must emphasize these mechanical phases of their work. The facts, however, are these: First, the use of standardized tests for teacher rating should be postponed, for the most part, until we have well-perfected tests in a greater variety of subjects. The tests should be used at present for pupil diagnosis, pupil motivation, pupil rating in particular subjects, for the trial of teaching methods, and so on, but not, except in a supplementary way, for teacher rating. Second, announced standards in the tool subjects should be thought of not only as goals to be reached, but also as limits not to be exceeded. In cities in which standardized tests are extensively used, bulletins are often issued informing the teachers that they are up to standard in the mechanical aspects of their work and that they should now devote time to other things. Third, the standardized *practice* tests enable a child to learn the tool subjects so much more quickly and thoroughly that more and better work is consequently made possible in the content and appreciation subjects. Early mastery of the Three R's allows the teacher to devote more and not less time to art and music and literature, and equip the pupil to do better work than before in geography and history and higher arithmetic, where these

tools will be of the greatest assistance. This is an effect that lasts through the upper grades and through high school and college. Fourth, the emphasis which tests give to the mechanical aspects of school work is at worst only temporary. Tests for the less tangible but perhaps more important factors in education are being rapidly developed.

Are the tests really constructed in a scientific manner? Every well-standardized educational scale represents the work of a specialist in a given field for months and sometimes for years. It is made up of exercises or problems which have been selected with the greatest skill and care to perform the difficult task of testing one particular ability of the pupil and only that one; which have been submitted to thousands of children to determine their real difficulty; which are prefaced by clear and brief directions found by trial to be easily comprehended by children; and which can be scored quickly and objectively. These scales are proved by ingenious experiment actually to measure what they purport to measure, and to measure it reliably or consistently. Before they are offered to others, they have been submitted to more tests than most of us would ever think of. Such scales are commonly placed on the market and sold for little more than printing costs; so the effort that has been put into their construction ordinarily receives no direct money return whatever.

For the teacher to take advantage of the results of such work and to utilize standardized tests in her classroom seems only common sense.

The Tests in Arithmetic

The Courtis Standard Research Tests, Series B, are among the oldest and most widely used of the arithmetic scales. They are probably used with more than a million school children each year. They consist of four tests, one each in addition, subtraction, multiplication, and division, of integers. Each test is printed on a separate page, and

the time allowances are, respectively, 8, 4, 6, and 8 minutes. The pupils are told that they are not expected to finish all the examples, but they are asked to work as rapidly and accurately as possible. There are two scores: the number of examples attempted, as a measurement of rate; and the number right, as a measurement of accuracy. The addition test consists of 24 examples, each 3 figures in width and 9 in height; the subtraction test consists also of 24 examples, each of these being 8 or 9 figures in width; the multiplication test contains 25 examples having 2 figures in the multiplier and 4 in the multiplicand; and the division test contains 24 examples having 2 figures in the divisor and 4 or 5 in the dividend.

In such a test there is obviously no opportunity for diagnosis; it is valuable mainly for comparison. It may be used for this purpose in any grade from the fourth through the eighth. The standards for each grade are based on the number of examples which pupils of that grade in different cities have been found able to solve.

The Woody Arithmetic Scale (for grades 2-8), also one of the earlier scales, is, on the other hand, of the diagnostic type. It consists of four scales printed on separate sheets, one for each of the fundamental operations. It goes beyond integers to include a few examples in fractions. Instead of having all the examples in one group to be of equal difficulty, as in the Courtis scale, it arranges the examples on each sheet in the order of their difficulty. Speed is, then, not measured; the pupil is allowed to make his way as far down the scale as he is able to go. The Woody scale is diagnostic in the sense that it includes in the scale for each operation examples of a great many kinds. By looking over the sheet the teacher can consequently tell, for instance, which variety of addition problems the pupil can work and which he cannot. But the examples are not arranged according to type—only according to difficulty—and the ability tested

by each example must be decided on by the teacher herself. Furthermore, there is, in certain editions of the scale, only one example to represent each type. One example is not enough to test a pupil's ability to work problems of a certain type.

The Cleveland Survey Arithmetic Tests (grades 3-8), which were designed at the time of the survey of the schools of Cleveland, Ohio, to give a more detailed appraisal of pupil ability than could be secured by the Courtis tests, consist of 15 tests of gradually increasing complexity, each test made up of a large number of examples of the same kind. Thus we have an attempt to make an analysis of arithmetic problems into different types, and a thorough test of the pupil's ability to deal with each type. The problems of the tests are arranged on a spiral principle—each operation recurring several times in a different and more difficult form. Thus, addition first appears as "Set A," which consists of 65 examples, each made up of two figures to be added together; it reappears as "Set E," in which there are 16 examples, each of five figures arranged in single column; then as "Set J," in which there are 14 examples of thirteen figures each, each example still a single column; and as "Set M," in which there are 12 examples, each five figures high and four columns wide. Thus it is seen that for addition there is tested successively knowledge of the tables, ability to add to a partial sum, attention span or the length of time the child can keep up this process and, finally, the ability to carry from column to column. Tests to cover different kinds of examples in subtraction, multiplication and division are arranged on the same spiral principle. In this way the teacher can discover which varieties of examples a pupil can and cannot work, and is able to give him help accordingly. A convenient feature is that a pupil's strong and weak points can be seen from a study of the scores arranged together on the cover of the book-

let, without leafing through the different pages. This test is diagnostic in a very helpful way. It can also be used for comparison, as it has standards based on surveys of several large cities.

The Monroe Diagnostic Tests in Arithmetic similarly present in a spiral form groups of problems of various types. They are different in that they are more complete and are printed in four separate leaflets. The first leaflet, Test I, is a series of very simple examples in addition, subtraction, multiplication, and division of integers, for use in the lower grades; Test II is more difficult examples of these operations with integers; Test III is a series of five groups of examples involving different "cases" in the four fundamental operations with common fractions; Test IV similarly involves the simple operations with decimal fractions.

These four measuring instruments in arithmetic show the process of development in this field. The Courtis test, one of the earliest, was a blanket test allowing no diagnosis. The early Woody test was a diagnostic test, but it did not attempt to analyze arithmetical abilities into types, nor to group problems along such lines. The newer Cleveland Survey test has succeeded in making such an analysis and grouping. The still newer Monroe test has retained this grouping, has extended the scale to include thorough tests of common and decimal fractions, and has broken the scale up into parts, so that one needs to make an expenditure only for the part directly adapted to the age of one's pupils.

These four may be regarded as typical of the "research" tests in arithmetic, whose function is to discover how well the pupils can perform the fundamental operations. In addition to these, there are the "practice" tests, whose function is to assist pupils to perfect themselves in these operations. Among the latter are the Thompson Minimum Essentials in Arithmetic (1908), the Studebaker

Economy Practice Exercises in Arithmetic (1916), and the Wildeman Practice Tests in Fractions (1922). The Courtis material consists of a series of about forty practice lessons printed on 5x8-inch stiff cards, arranged in order of complexity. Each pupil is provided with a tablet of the same size, made up of sheets of transparent paper. When the pupil inserts a card beneath a sheet of this thin paper, the figures on the card show through, and the pupil writes his answers on the tissue paper. Thus the card itself is preserved intact. On the back of the card the answers are printed, so that when the inverted card is placed in the tablet the answers appear just below the pupils' answers, and in this manner the practice work can be corrected by the pupil himself. The pupil thus proceeds from one card to another as rapidly as he masters the given type of problem. At intervals *test* cards are inserted—cards without answers on the back. In case of these the tissue sheet is torn out and corrected by the teacher, who lays it over an answer sheet in her Manual of Directions. Each child proceeds at his own rate and is always working on the kind of example which he needs next to master. He keeps his own record of progress upon graph sheets found in his tablet, and thus his progress and his objective are always obvious and definite.

The principle involved in such practice exercises is mastery of the educational skills by individual work. It means abandoning the idea that skills can be taught by mass instruction. It means a partial return to the individual teaching prevailing in the days of our grandparents, but a return after all to a quite different level. For the newer individual teaching utilizes skillfully graded practice material. It not only allows each pupil to proceed at his own rate, but assists him by these graded exercises to proceed systematically and therefore to master the essentials rapidly and thoroughly.

The Courtis exercises, taken above as illustrative, have brought steady improvement in the fundamentals in several different cities. They have proved that allowing each pupil to practice daily on the kind of problem which he (and perhaps no one else in the class) has reached in his development on that particular day, does result in rapid mastery of the basic skills in arithmetic. Others of the practice tests, however, are sometimes considered less difficult to administer. Thus, the Plymouth and the Studebaker practice exercises avoid the necessity of using thin paper (which under small sweaty hands exhibits a regrettable tendency to curl up and become unmanageable) by printing the examples on cards which have cut-out portions through which the answers can be written on ordinary paper. The Wildeman exercises use still another scheme, printing the practice problems in a graded series in a small leaflet, which the pupil utilizes by laying his paper just below the problem he is to solve. The paper can then be folded down and another answer written, and so on indefinitely. A plan very similar to this has proved successful in schemes for individual teaching worked out in the schools of Winnetka, Illinois.

The teacher should by all means secure descriptive literature and samples of these practice tests, and of others now available in handwriting and reading, and should plan to change over her teaching of the Three R's to an individual basis as rapidly as possible. In this connection, she should read the accounts of the experiments and results in the Winnetka schools, published in the *Elementary School Journal,* September, 1920, and in the *Journal of Educational Research,* March, 1922. All promotions in that city for two years have been based on individual work. Each pupil is allowed to go forward at the rate best suited to his ability. But, of course, practice tests will also serve as teaching devices in schools not using a scheme of individual promotions.

Turning now from the tests on the fundamental operations to the tests involving reasoning, we find problems of the type which are written out in words, such as, "How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?" In these problems the operation to be performed is not indicated, and the ability tested is just the ability to decide which operation to perform. The oldest of these tests is the Stone Reasoning Test in Arithmetic, first published in 1908. It consists of twelve problems about school happenings or other interesting affairs, arranged in order of increasing difficulty. It has been used in several school surveys, and therefore has satisfactory standards. It does not attempt to classify arithmetical problems into types.

The Monroe Standardized Reasoning Tests in Arithmetic (1918) have made this classification. Just as the Cleveland Survey tests went beyond the Woody by attempting to analyze the ability to add or subtract into several distinct abilities and to devise a separate test for each, so the Monroe tests go beyond the Stone by attempting to analyze reasoning problems into types and to test these separately. The Monroe scale consists of three tests, issued separately. Test I, for the fourth and fifth grades, is built around simple operations with integers; Test II, for the sixth and seventh grades, involves fractions; and Test III, for the eighth grade, involves percentage. The problems use the form of language statement found to occur most commonly in eight widely-used textbooks. A pupil may be given three scores: one for rate, one for use of the correct principle in his solution, and one for getting the correct answer.

The Buckingham Scale for Problems in Arithmetic (1919) is quite similar in type. The first division is for grades 3 and 4, the second for grades 5 and 6, and the third for grades 7 and 8. The problems are arranged in order of statistically-determined difficulty, and the

pupil's score is the value of the last problem he succeeds in solving. No time limit is used.

To find whether her pupils are up to standard in arithmetic, the teacher should give one of the research tests on the fundamental operation and one of the reasoning tests. If her pupils are below standard in the operations, she should give them practice in this work about ten minutes a day with one of the sets of practice tests. Enabling the pupil to have definite objectives and to see his daily accomplishment in comparison with those of other pupils of his own age, is found wonderfully effective in securing improvement. Psychologists state that the most influential single factor in learning is the purpose to learn. Mere repetition of an exercise—unmotivated drill work—does not bring mastery of that exercise. But when a pupil tries hard because of a definite purpose, his learning is rapid. Any adult can verify this from his own experience in learning typewriting or piano fingering or golf or a foreign language. One-half hour of work under concentrated attention is worth three hours of work of a dilatory kind. People fix their attention on their work when they have a definite purpose in mind. One of the best ways to assist pupils to definite purposes is to give them definite daily objectives. The practice tests are so arranged as to do this.

If her class is shown by the research tests to be up to standard in the fundamentals and below standard in the reasoning problems, the teacher should not spend time on the practice tests. She can then with confidence omit repetitional work with the operations and can devote extra time to instruction in reasoning problems, until a second test shows that the class has attained standard proficiency in those also.

The Tests in Reading

Silent reading ability is the ability to get the thought

from what one is reading, while oral reading ability is the ability to transmit this thought to others. The two are tested separately.

The simplest of the silent reading tests are the vocabulary tests. These consist of lists of words whose meaning the child is to indicate. In the Thorndike Visual Vocabulary Scales (1914) the words are arranged in lines of ten words each, all the words in one line being of equal difficulty. The child is directed to write the letter F under every word that means a flower, the letter A under every word that means an animal, etc. The words get harder as one works down the page, and the pupil's score is the score value of the last line in which he gets at least eight of the ten words correct. Since all the answers are either right or wrong, this little test is quite objective and is very easy to give and score. It will be useful in showing the teacher the extent of the vocabulary possessed by a class she is taking charge of, or in showing the source of certain pupils' difficulty in reading—whether the cause is failure to understand the meaning of the words, or some other deficiency.

A second type of silent reading test measures the pupil's ability to understand reading material expressed in sentences and paragraphs. One of the earliest and most widely used is the Thorndike Scale Alpha (1916). The scale consists of a series of paragraphs, each followed by questions to be answered by the pupil. The paragraphs are arranged in order of difficulty, and the range is such that the lower part of the scale can be understood by very young children, while the upper part requires rather close attention even from an adult. No time limit is used and the child is allowed to go as far as he can. Sample paragraphs follow:

Set II. Difficulty 5.25

Read this and then write the answers. Read it again if you need to.

Long after the sun had set, Tom was still waiting for Jim and Dick to come. "If they do not come before nine o'clock," he said to himself, "I will go on to Boston alone." At half past eight they came, bringing two other boys with them. Tom was very glad to see them and gave each of them one of the apples he had left. They ate these and he ate one too. Then all went on down the road.

1. When did Dick and Jim come?_____
2. What did they do after eating the apples?_____
3. Who else came besides Jim and Dick?_____
4. How long did Tom say he would wait for them?_____

Set IV. Difficulty 7

Read this and then write the answers to 1, 2, 3, and 4. Read it again if you need to.

You need a coal range in winter for kitchen warmth and for continuous hot-water supply, but in summer when you want a cool kitchen and less hot water, a gas range is better. The XYZ ovens are safe. In the end-ovens there is an extra set of burners for broiling.

1. What effect has the use of a gas range instead of a coal range upon the temperature of the kitchen?_____
2. For what purpose is the extra set of burners?_____
3. In what part of the stove are they situated?_____
4. During what part of the year is a gas range preferable?_____

It will be seen that some of these answers are to be written out in sentences. This requirement makes the scoring rather cumbersome, or else somewhat unreliable. To make it possible for everyone to score the papers with the same results, the author supplies a list of possible right and wrong answers. But to take the trouble to look these up is time-consuming, whereas to rely only on judgment as to whether an answer is correct or not is to get results that are not thoroughly objective. This scale has been widely used, however, and its standards are based on a large number of scores.

Another group of tests which have been widely used are the Kansas Silent Reading Tests (1916). Sample paragraphs follow:

No. 1
Value 1.0

The air near the ceiling of a room is warm, while that on the floor is cold. Two boys are in the room, James on the floor and Harry on a box eight feet high. Which boy has the warmer place?_____

No. 14
Value 4.9

A list of words is given below. One of them is needed to complete the thought in the following sentence: The roads became muddy when the snow _____. Do not put the missing word on the blank space left in the sentence, but put a cross below the word in the list that is next above the word needed in the sentence.

water
is
melted
snow

These tests have a time limit of five minutes, and the rate score depends on the number of paragraphs about which answers were made out in that time. The exercises in the test are very interesting for the pupils, and are very easy for the teacher to score. Some of the exercises, however, are very much like puzzles and others are like arithmetical problems, both of which require thought *about* what is read, rather than comprehension of what is read. The weakness of the scale, then, is that it does not always measure what it purports to measure. But the scoring is thoroughly objective and the standards are adequate.

A set of tests very similar to the Kansas scale, but free from some of its weaknesses, is the Monroe Standardized Silent Reading Paragraphs. These attempt to limit the question to what would be grasped by *understanding* the paragraph, rather than by adding one's thought to it. They have a time limit, and measure rate and comprehension separately. The rate score is the sum of the rate values given each paragraph, and the comprehension score the sum of the comprehension values of each paragraph. Test 1 is for grades 3, 4 and 5; Test 2, for grades 6, 7 and 8; Test 3, for grades 9, 10, 11 and 12. The paragraphs are arranged in order of difficulty, so that everyone

is able to make a score. Here are sample paragraphs from Test 1:

No. 2

Rate Value 7	The little Pilgrim girls carried their work boxes to the dame-schools and learned to sew and knit as well as to read and write. Where did the girls go with their work boxes? To the_____	Comprehension Value 1.3

No. 4

Rate Value 9	Hiawatha was a little Indian boy. He had no father and no mother. He lived with his grandmother, Nokomis. His home was in a wigwam. Draw a line under the word that tells whom Hiawatha lived with. Father, aunt, mother, uncle, sister, grandmother.	Comprehension Value 1.4

Another of the newer tests is the Courtis Silent Reading Test No. 2. This consists of a story of about five hundred words which is first presented as a whole. It secures the rate measurement by having the children read through the story for three minutes, while at the end of every thirty seconds, when the teacher says "Mark," each pupil draws a line around the word which he is reading at that time. The comprehension measurement is secured by presenting the story a second time broken up into paragraphs with questions after each paragraph. In order to make the scoring completely objective, all the questions are of a form that can be answered by *yes* or *no*. The effect of guessing, the pupil having an even chance to guess the answer correctly, is cut down by subtracting the number of wrong answers from the number right. The answers to the questions do not require the pupil to go beyond the material in the story. The test is adapted to grades 2-6. Here is a sample paragraph from the complete story called "The Kitten Who Played May-Queen":

When the day of the party came, Daddy planted a May-pole and Mother tied it with gay-colored ribbons. There were to be

games and dances on the grass and a delicious supper, with a basket full of flowers for every child.

1. Were the children to have anything to eat?_____
2. Were they going to play on the grass?_____
3. Were they going into the house to dance?_____
4. Were the baskets to be full of flowers?_____
5. Was it Daddy who tied the ribbons to the pole?_____

The Burgess Pictorial Supplement Scale (1921) is built on quite a different plan. Mrs. Burgess criticizes the previous tests in reading as in some cases measuring other abilities besides reading ability; in others, being made up of such a variety of exercises as to make it difficult to interpret the results; and in others, being hard to give and score. The Picture Supplement Scale is made up of exercises all of one type. A drawing is shown, and just below is a paragraph asking that something be done to the picture. The thing to be done is very simple, but the child cannot do it unless he understands the directions as they are given in the printed form. In connection with the sample paragraphs given below, one must use his imagination to supply the drawing that in each case appears just above the paragraph.

1. This naughty dog likes to steal bones. When he steals one he hides it where no other dog can find it. He has just stolen two bones, and you must take your pencil and make two short, straight lines to show where they are lying on the ground near the dog. Draw them as distinctly as you can and then go on.

2. This man is an Eskimo who lives in the far north where it is cold. There has just been a big storm, and all the ground is white with snow. The man has been walking and has made many footprints in it. With your pencil quickly make four of them in the snow just behind him.

19. When the road is rough the porter finds it hard to push this wheel chair. Draw a line to show where the road is. Be sure to make the line in front of the chair smooth so that the chair will roll along easily, but make the line in back of it uneven because up to this time the path has been rough.

In this test there are no puzzles or catches, and the

wording is simple throughout. The vocabulary is taken from the commonest words in the English language, as revealed in previous studies of correspondence and newspaper articles. The number of ideas in one paragraph has been reduced as far as possible and organized around one central idea. All the paragraphs are carefully constructed to be alike in each of these characteristics. All factors which would modify the score without being strictly a part of reading ability have been supposedly ruled out—such factors as demand for special imagination, or for ability to remember or to reason. Others, which could not be eliminated, have been held constant—difficulty of action demanded, vocabulary difficulty, sentence structure, uniformity of print, interesting character of paragraph, etc.—so that only one factor is to vary and be measured, namely, the amount a child can read and understand in a given time. Although six different scales for measuring silent reading were completed, printed, and tried out in 23 school systems before this scale was perfected, the Picture Supplement Scale was finally retained as that best meeting these requirements. Some of the other scales were of the type in which the exercises appear in order of increasing difficulty and the child is given no time limit, but is allowed to go as far as he can. These were finally rejected, for the author believes that rate of reading is an important factor in reading ability. Of several typists who are equally accurate, the one who turns out the most pages per hour is the best worker, and of several newspaper reporters who write equally accurate and interesting stories, the one who gets his copy to the editor's desk with the greater speed is considered best. Similarly, the child who by endless rereading and rechecking can get a piece of work right is not the best student. A time limit of five minutes is therefore used in these tests.

Summary. In the reading tests here described we see a development from mere word lists and from rather

cumbersome and semi-subjective tests of connected discourse toward tests which measure the complex reading ability in a simple and objective manner and without a marked modification of the score by other abilities. Comprehension has been measured, in these and other reading tests, by asking the meaning of separate words, by asking a reproduction of the story (Starch test), by asking questions about the story, by a combination of these two (Gray test), and by asking the pupil to carry out certain directions. Rate (sometimes disregarded) has been measured by counting the number of exercises completed in a limited time, by counting the words in the exercises completed, and by counting the number of words read every thirty seconds when no pauses were made for writing answers. By availing herself of the results of thousands of hours of intensive work by experts, the teacher is now able to get a very accurate measurement of the rate and the completeness of her pupils' comprehension of what they read. And silent reading is probably the most important study in the curriculum, because it is the key to almost all the other studies.

Oral reading can now also be measured. The simplest tests here, as in silent reading, are word lists; but in this case, of course, the ability tested is not the ability to give the meaning of the word, but to pronounce it. In the Haggerty Visual Vocabulary Tests, for example, the words are arranged in groups of uniform difficulty of pronunciation, and the child's score is the value of the last group in which he can pronounce correctly 80 per cent of the words.

A more complete measurement is secured by using the Gray Oral Reading Test, which consists of about a dozen successive paragraphs in which the words grow harder and harder to pronounce. The lower end of the scale can therefore be used with very young children, while the upper end of the scale is sufficiently difficult to give pause

even to a mature pupil. This can be seen from the two following widely separated selections. The type used for young children is of course very much larger than that here shown.

1. A boy had a dog.
   The dog ran into the woods.
   The boy ran after the dog.
   He wanted the dog to go home.
   But the dog would not go home.
   The little boy said, "I cannot go home without my dog."
   Then the boy began to cry.

II. The hypotheses concerning physical phenomena formulated by the early philosophers proved to be inconsistent and in general not universally applicable. Before relatively accurate principles could be established, physicists, mathematicians, and statisticians had to combine forces and work arduously.

While the pupil reads a paragraph aloud, the teacher marks on her own copy, by a definite system of scoring, errors of six types: gross errors, minor errors, omissions, substitutions, insertions, and repetitions. The directions furnished with the test define and illustrate these errors and the method of indicating them and deducting for them in very specific terms, so that a number of teachers scoring the same pupil would give him the same score. This test has been given to a large enough number of children to furnish quite reliable standards by school grades. It is necessarily an individual test, which makes its use rather laborious in comparison with the group tests we have been describing. For the measurement of the reading ability of selected individual pupils whose oral reading needs special study, however, it will be found very valuable.

The Tests in Spelling

The best-known scale for measuring ability to spell is the Ayres Spelling Scale, which is made up of the one thousand commonest words in the English language arranged in twenty-six columns, each column made up of words found by many thousand children about equally

difficult to spell. Which words of the English language occur most frequently was discovered by studying letters, newspapers, and standard literature, in an amount aggregating nearly 400,000 words.

At the top of each column is shown the per cent of correct spellings to be expected in each grade. These are based on the spellings of about 70,000 children. The scale, on account of its content, has been used so extensively as curriculum material in spelling that its value as a measuring device may presently disappear, for the early median scores will not continue to represent average attainment. But when this happens it will mean that the scale has been effective in insuring in the average American child the ability to spell correctly the words he will most commonly need to know how to spell.

One trouble the teacher finds in using this scale is that there are hardly enough words in a column (words of equal difficulty) to serve to make up a test. To meet this situation we now have the Buckingham Extension of the Ayres Scales, in which the original list is enlarged to about 1,500 words. The new words were selected, however, not on the basis of their frequency of actual use but on the basis of the frequency of their occurrence in standard spelling books. The Iowa Spelling Scale, which was built up, in a manner similar to that used by Ayres, from the vocabulary found in the correspondence of the citizens of Iowa, contains nearly 3,000 words.

Presenting words for spelling by pronouncing them one at a time is not, of course, presenting them in a "natural situation." The attention of the pupil is fixed on the word itself, while in actual writing his attention is fixed on the meaning he is trying to express or on the relation of the words to each other, and the spelling of the word is in the margin of one's attention. To test spelling in a situation more nearly resembling that in which the ability to spell will actually function, Monroe has devised a

Timed-Sentence Spelling Test. In this the words to be spelled are embedded in sentences, which are dictated at a regular rate for copying. None of the test words are placed near the end of the sentences, and in order to allow for differences in rate of writing the pupils are told that if they have not finished a sentence before another is dictated they are to leave it and begin the new sentence. Since this is one of the newer devices, this test does not have standards as good as those of the Ayres scale. For use in experiments with devices for teaching spelling, however, where it is only comparison of earlier with later achievements that is needed, the lack of standards is not serious.

Tests in Punctuation and Grammar

The punctuation scales are made up of sentences printed without punctuation marks. The punctuation is to be inserted by the pupil. Such tests are used quite extensively in business schools and in business firms, as well as in the public elementary school. They are very simple in structure, and are easily adapted to teaching as well as to testing.

The grammar tests present sentences which are grammatically incorrect and are to be corrected by the pupil. The Charters test, for example, has separate scales for pronouns and for verbs. Like the punctuation scales, these are not difficult to understand or to use.

The Composition Scales

The earlier scales for measuring merit in English composition were blanket scales, measuring all types of such ability together. The first, the Hillegas Scale, consisted of ten brief sample compositions arranged in order of merit, without any description of the good and bad points supposedly present in each sample, and without distinction between the various forms of writing. The Harvard-

Newton Scales, published later, contains four separate scales, one each for description, exposition, narration, and argumentation. A short analysis is printed below each selection, showing its points of strength and weakness. The Willing Scale, which has been used in a survey of the schools of Denver and other cities, consists of eight brief compositions on the theme: An Exciting Experience. The compositions are rated separately for "story value" and for "form value." The Lewis Scales for Special Types of English Composition measure ability in letter writing. There are scales for judging order letters, letters of application, social letters of the narrative type, and social letters of the problematic type. Letter writing is a form of composition for which scales are especially useful.

In using a modern composition scale, the teacher secures from her class compositions about one of a list of suggested subjects, written under carefully described conditions; and she then compares each of these with the standard samples shown on the scale. The teacher should remember that to estimate a composition accurately by comparing it with the scale requires a good deal of skill. The composition scales, as contrasted for example with the arithmetic or spelling tests, are for this reason comparatively hard to use, and measurement by them should not be attempted without a good deal of preliminary training. Such training can be secured by practice in rating compositions which have already been rated by an expert. A collection of such exercises for practice has been published by Thorndike. It is called "English Composition, 150 Specimens Arranged for Use in Psychological and Educational Experiment," and can be procured from Teachers College, Columbia University, New York.

The Handwriting Scales

Handwriting is like English composition in that it is

measured by comparing samples of pupil work with standard samples arranged in order of merit on a scale. The handwriting scales are of two types: those for comparison and those for diagnosis.

The Ayres scale, one of those most commonly used for comparing one group of pupils with another or with standard achievement, is made up of examples of penmanship which are arranged in order of legibility only. Legibility in these selections was determined by timed readings, and the order is therefore completely objective. In the "Gettysburg Edition" of this scale the wording of all the samples is the same, being the first few sentences of Lincoln's Gettysburg Address. When the pupils to be tested also write from this address, comparison of their work with the scale is made more accurate. Standards for this scale are very dependable.

The Thorndike scale, which is also extensively used for comparison, is made up of fourteen examples of penmanship arranged according to a composite of three criteria: beauty, legibility, and general merit. Their order was determined not objectively but by the consensus of opinion of a considerable number of handwriting experts, teachers and supervisors of handwriting. Conversion tables have been worked out which show the value of each of the fourteen Thorndike samples in terms of the eight Ayres samples, or vice versa. By this means one using either of these scales is able to avail himself of the standards of both.

The Freeman scale, on the other hand, is designed for diagnosis. It really consists of five scales printed on one sheet. These five measure, respectively, uniformity of slant, uniformity of alignment, quality of line, letter formation, and spacing. A pupil whose score is low on the Ayres or Thorndike scales, or whose writing is otherwise known to be poor for his grade, can be 'diagnosed' by

means of this scale and the exact defect in his writing discovered.

The Gray Score Card, modeled perhaps after score cards for judging livestock, is simply a list of the principal characteristics of handwriting, such as alignment, size, slant, etc., with a value opposite each from which deductions can be made. Such a card when filled out shows the strong and weak points of the pupil's writing, and indicates to pupil and parent the points in which he needs to improve. By making a 'diagnosis' of a pupil's writing as a physician does of his physical condition, the teacher can direct the pupil to concentrate his practice on certain points of technique and, by referring to the diagnostic card later on, can tell whether the practice has been effective. The purpose of the card is the same as that of the Freeman scale, the difference being that the Freeman scale presents samples of writing with which the pupil's work can be compared while the card gives only a list of the names of the qualities measured.

In giving tests in handwriting the teacher cannot secure a measurement of rate by having the pupils copy material from print, for in that case the rate of writing is confused with the rate of reading; nor by having the pupils write from dictation, for then the rate of writing is governed by the rate of dictating. The children must write a few lines from memory. Whatever selection is taken for this purpose—whether it be Mother Goose Rhymes or the Gettysburg Address—should be reviewed until it is freshly in mind before the test begins. The pupil should then write at his regular rate for the period of the test. A key copy of the selection can be prepared by the teacher for help in scoring the results, annotated to show number of words occurring up to any given point. The number of words written by each pupil in the time allowed can then be seen by a glance at this key. This gives the score for rate. The score for quality is secured by finding

the standard sample which is most like the writing of the given pupil.

The handwriting scales, like the composition scales, will not yield reliable scores until the tester has had considerable practice in making the comparisons involved. Such practice can be facilitated, as in the case of English composition, by securing a large number of writing samples the merit of which has been determined and recorded opposite the number of the sample in a key list. Such a group of rated samples has been issued by Thorndike in a booklet entitled, "Teachers' Estimates of the Quality of Specimens of Handwriting," procurable from Columbia University.

Other Achievement Tests

The school subjects discussed above are those in which tests and scales are now best developed. But there are also standardized tests in history, geography, drawing, music, journalism, physical training, manual training, home economics, commercial subjects, algebra, geometry, general mathematics, general science, physics, chemistry, biology, Latin, German, French and Spanish. Samples of these tests may be secured from the publisher named in the Bibliography referred to on page 36.

"Home-Made" Objective Tests

A teacher who wishes to secure a more accurate measurement of her pupils, not so much for the purpose of comparing them with pupils elsewhere as for the purpose of comparing them with each other or with her previous classes, can do so without depending on the published standard tests. She can make a test of her own. To do this she should construct a series of simple and unambiguous statements about the material recently covered in class. The following examples, taken from courses in Education, will show what is meant:

True-False Test

Directions: Place before each statement the word *true* or the word *false.*

_____11. The socialized recitation is better adapted to dull pupils than to bright ones.

_____19. Scientific bases for the curriculum have been more carefully worked out in spelling than in geography.

Best-Reasons Test

Directions: In each case, place a cross before the best reason.

1. If the project method is better than the older methods of teaching, that is because:

It develops the pupil's originality.
It is being advocated by most colleges of education.
It can be more quickly mastered than the older methods.
It makes sure that all pupils cover the same ground.

2. The reasoning tests in arithmetic are not so widely used as the tests on the fundamental operations because:

It takes more time to give the reasoning tests.
Reasoning ability is comparatively hard to measure.
The designers of the operations tests are men of more prestige.
The reasoning tests have only recently been devised.

Multiple-Answer Test

Directions: Indicate the correct answer by underlining one word or phrase in each parenthesis.

3. The first scientific educational scales were worked out in (history, geography, penmanship, geometry).

10. The most reliable test for the measurement of the intelligence of children is the (Voelker, Liao, Binet, Thurstone).

This kind of examination, when used at the end of a semester, has many advantages over the traditional examination. For one thing, it covers a great deal more ground. A class can easily mark from fifty to seventy-five of these statements and have time left to score the papers during the same class-hour. The large number of statements means a more thorough review by the pupil, less luck in grades from getting questions on just the material that happened to be reviewed, and a more thorough test by the teacher. Again, the score is independent of the teacher's subjective standards of scholarship. The

score depends only on the number of right and wrong answers, and will be the same no matter who grades the papers. This means a fairer rating of the pupils. It also means a better relationship between teacher and pupil, for the teacher instead of being a judge, a person standing perhaps between the student and his diploma, can take the role of a guide and helper who assists the pupil to secure that mastery of subject-matter which will be indicated by a high score on this objective and impartial examination. Again, such an examination (when once constructed) is time-economy and spares the teacher the worst drudgery in teaching—the reading of a large number of answers monotonously alike. This type of examination also saves the pupil from the drudgery of writing out long answers. In the author's classes, and in others which have reported using the test, the students after trying it have always voted almost unanimously in favor of this kind of an examination. Furthermore, the test can be used as an excellent teaching device, for if it is given so as to allow another hour for discussion of the statements, it will serve as a very good outline of the course to date; and the facts in it are likely to be remembered if they are talked over after being presented to the concentrated attention characteristic of the examination hour. Another advantage of such a test is that if the test papers are always collected after the discussion, alternate forms of the test can be used year after year and standards of attainment can be gradually built up. This allows one to grade a given pupil by comparing him not only with the other members of his class, but with all previous classes which have taken the test.

Certain disadvantages in the use of such an examination will, to be sure, be found. There is, of course, a chance to get the answers correct, especially on the true-false type, by lucky guessing. This is cut down by the scoring scheme used for the true-false type; namely, sub-

tracting the number of wrong answers from the number right. This eliminates guessing, in case there is a large number of statements, as can be seen from the following formulae. If x is the number of statements for which the pupil knows the answer; y, the number he guesses right; and z, the number he guesses wrong; then the total number of right answers is x plus y, and the final score is x plus y minus z. But since he has an equal chance to guess a statement right or wrong, the number guessed right will be approximately equal to the number guessed wrong; that is, y is equal to z. Then in the formula, *score* equals x plus y minus z, y and z will cancel each other, leaving the score equal to x, or the number known, as it should be. This demonstration, however, is based on the assumption that the number guessed correctly and incorrectly will be about the same, which is true only in case a considerable number of statements can be guessed at, just as it will be found that a coin when tossed will show the same number of heads and tails only when a large number of tosses are made. For these reasons, a true-false test ought to contain a hundred or more statements if its results are to be used as the basis of important school marks or grades. These limitations are not found in the best-reasons and multiple answer types of examination (when some four or five answers for each question are shown), and it is therefore such types that are particularly recommended to the teacher.

Another disadvantage of these examinations is that they do not test the pupil's power to *organize* his thought This drawback can be dealt with, however, by supplementing the objective examination with short essay papers written at home. These probably give a better measure of the pupil's power to organize, and better practice in organizing, than do the answers hastily put together in the stress and hurry of an examination hour. The use of such a plan will involve the teacher again in reading

student papers, but it will be a type of reading not half so tiresome. For the home essays can be written on topics chosen from a considerable number of alternatives and can be prepared with more care, and they will therefore furnish a variety of reading that is often really very interesting.

Of course, these objective examinations made by the teacher herself are less accurate measuring instruments than the standardized tests, because they are made up of statements which, while not known to be equal in difficulty, are usually given equal credit. And they are less useful diagnostic instruments, because the statements are not ordinarily arranged according to the *type* of information implied in correct answers. Their scores should not, therefore, be considered to be as dependable as those on real standardized tests. But we should also remember that, in spite of these limitations, these examinations have been found by actual trial to be at any rate more accurate than the traditional examination. In fact, the traditional examination suffers from exactly these same limitations, as well, it might be added, as from many others.

In giving such objective examinations, it is best to have the statements multigraphed. A card having on it the pupil's name and a key number can be given to each member of the class, and the number instead of the name can be placed on the paper. When the papers are redistributed to be scored as the teacher reads the answers, this anonymous character of the papers promotes honesty in the scoring. A further check can be secured by having each pupil write on each paper he scores: "Scored by ————," signing his name. By devices of this kind, pupil-scoring can be made reliable, and the time the teacher ordinarily spends in reading test papers can be saved for something more profitable.

If it is not practicable to have the examination multigraphed, the statements may be written on the blackboard

in advance, or they may be read aloud by the teacher. In such a plan the pupil may take down only the *number* of the statement, and indicate his answer by writing "true" or "false" beside it, or in the case of the best-reasons test by writing "second reason" or the like, or in the case of the multiple-answer test by writing the single word needed for the answer.

That such examinations actually give dependable results has been proved by comparing their scores with other records of scholarship, such as grades on traditional examinations in the same class, or the average grades at the end of the previous semester. By the latter criterion, the new objective examination stands higher than the traditional examination. Successive ratings of the same class by an objective examination have also been found to agree better with each other and with home-essay ratings than the traditional examinations do with each other or with home essays.*

Fully standardized tests in sufficient numbers to cover all parts of the curriculum are probably a long way in the future. In the meantime the teacher can give a very much fairer rating to the work of her pupils, can establish better relations with them, and can relieve both herself and her pupils of a great deal of drudgery, by using some form of the "home-made" objective examinations here described. The only hindrance to putting such a plan into general effect is the time it takes a teacher to collect and phrase clearly and briefly the large number of statements which it requires. With the older children, however, such formulations can be made out by members of the class as home assignments. This scheme is an excellent one for giving practice in discovering the most important points in a chapter, and for well-motivated practice, in a "real situ-

* For a statistical presentation of this evidence, see Gates, *The True-False Test as a Measure of Achievement in College Courses, Journal of Educational Psychology,* May, 1921. Or, Wood, *Measurement of a College Work, Educational Administration and Supervision,* September, 1921.

ation," in effective use of terse and unambiguous English.

The true-false and best-reasons and multiple-answer examinations should be thought of as the intermediary between the old-fashioned examinations whose ineffectiveness was discussed in Chapter I, and the tests which can really be called standardized tests. The objective examination made by the teacher will serve to bridge the gap from the one to the other.

What to Do

The reader who wishes to become familiar with the standardized tests which have so far been worked out for her own subject or grade should now write to the principal publishing houses, named on page 37, and from their price lists should order samples of the tests which seem best fitted to her needs. She should also secure the Bibliography of Tests mentioned on page 36, and, when possible, read the cited reference to a more complete account of the test by its author. By means of these, and by means of the discussions of the tests in some of the longer textbooks cited at the end of this chapter, she will be able to decide which of the tests she wishes to secure.

Selecting an Achievement Test

In making a selection among the numerous tests and scales which are now available, the teacher should keep in mind the exact purpose of the experiment in hand. If the purpose is to compare a room or school with average attainment, then she should choose a test which has been widely used and which bases its standards upon a large number of cases. If the purpose is to measure pupil progress over a series of weeks or months, then she should choose a test which has a considerable number of "forms" (duplicate editions), in order that the score at the time of the second testing may not be influenced by the pupils' remembering any of the previous answers. If the purpose is diagnosis, then she should choose a test which is finely

divided and measures the various related abilities separately.

The teacher should also make sure, of course, that the selected test is adapted to the age of the children she is teaching. The time used in administering the test should be considered, in order that it may not exceed the length of the class period, unless one wishes to extend the period or to break up the test into parts.

In case of a scale, such as the handwriting scales, one should also consider the time necessary for learning to use the scale accurately. Not the least important consideration is money cost, and this may sometimes be kept down by selecting a test which is designed for a particular grade rather than for a series of grades. Having an uncopyrighted test multigraphed is not advisable if one is to compare her class with standards, since the change in the appearance and the arrangement of the problems will affect the scores. Almost no outlay of money will be necessary for measurements by such scales as those in handwriting or English composition, for these require only one or two copies of the scale for the entire room. But such scales are probably the most expensive of all from the point of view of time-cost.

Instructions for giving and scoring the tests are usually enclosed in the package of test blanks when it is sent out by the publisher. These instructions, as has been said before, should be followed to the letter. Methods for working up the scores, in ways that will show what they reveal about a class, are explained in the next chapter.

The teacher who wishes a fuller discussion and description of achievement tests than is possible in so brief a sketch as this, should consult some of the following textbooks, which are arranged according to date of publication:

Monroe, DeVoss, and Kelly: *Educational Tests and Measurements.* 1917. (Houghton Mifflin Company.) Describes a considerable number of the tests in use in each

of the principal school studies in 1917, and briefly explains some practical statistical methods for treating results.

Monroe: *Measuring the Results of Teaching.* 1918. (Houghton Mifflin Company.) Gives more emphasis than its predecessor to the remedial instruction which should follow application of the tests.

Wilson and Hoke: *How to Measure.* 1920. (Macmillan Company.) An elementary handbook for those giving tests.

McCall: *How to Measure in Education.* 1922. (Macmillan Company.) An advanced textbook, primarily of interest to school supervisors and administrators. Explains how to use educational tests, how to construct and standardize them, and how to use statistical methods. Technical and difficult.

Pressey: *Introduction to the Use of Standard Tests.* 1922. (World Book Company.) An elementary outline presenting in simple language much practical information.

CHAPTER V

PUTTING MEANING INTO SCORES

Tables

The large mass of scores resulting from a test series have very little meaning for the teacher just learning to use measurements. She is not able at first to see what they reveal about her class. She can interpret her results only by learning something of those ingenious methods for handling mass data which were developed for this purpose in economics and biology, and which have more recently proved so influential in psychology. When once understood, the computations involved are really very simple.

The first thing to do in finding the meaning of a set of scores is to put the scores in order. The test papers may be sorted so as to have the best paper at the top of the stack, the next best paper just below it, and so on down. It will then be desirable to make a record of the scores, in this way:

Table I

Pupil's Name	Score
Henry Jones	27
Mary Brown	25
Thomas Smith	24
Charles Green	24
Etc.	Etc.

By a glance at such a completed table one can see the general distribution of the scores—what the highest and lowest scores are, and what are the scores in the middle of the table, which the average child of the class is able to make. It will usually happen that a number of pupils

will earn the same score. A shorter and more convenient table can then be made in this way:

Table II

Score	Number of Pupils Making the Score (Frequency)
58	1
57	2
56	2
55	3
54	5
53	8
52	5
51	4
50	4
49	3
48	1
47	1

This is called a Frequency Table. It shows the frequency with which each score occurs.

When the scores spread over a very wide range and there are a considerable number of them, it is convenient to arrange them in groups, in some such way as this:

Table III

Score	Frequency
140-149	2
130-139	4
120-129	7
110-119	6
100-109	7
90- 99	10
80- 89	12
70- 79	11
60- 69	8
50- 59	6
40- 49	4
30- 39	2
20- 29	1

This table says that there were 2 pupils who made scores between 140 and 149 inclusive, that there were 4 pupils who made scores between 130 and 139 inclusive, and so on. A very large number of scores can be recorded in a compact and intelligible form by this plan.

Of course if the range in the scores had not been so wide as here indicated, a smaller interval could have been used for the score column, thus:

Score
115-119
110-114
105-109
100-104
etc.

Here we are using intervals of five instead of intervals of ten. In the case of the set of scores represented by Table III, intervals of five would have made the table too long to be convenient. The table would have contained twenty-six divisions instead of thirteen,—an unnecessary fineness of distribution for most purposes. A good working rule is to so choose the intervals of the score column that the table will have some ten or twelve divisions. This can be done by subtracting the lowest from the highest score, pointing off one place (dividing by ten), and taking the nearest convenient unit. For example, if 22 and 146 are the low and high scores in the group represented in Table III, then $146 - 22 = 124$, and pointing off one place gives 12.4, from which the interval may be taken as 12, or as the nearest round number, 10.

The Median

In order to compare one class with another, or with a standard, we need some one number to stand for the class as a whole, to represent the 'central tendency' of the class. Our first thought would be to compute the 'average' of the scores. But there is a measure of the central

tendency of the class which is so easily found that it is in universal use among makers of standardized tests, and must therefore be used instead of the 'average' by those who would compare their results with standards. This measure is the median. The median may be thought of simply as the middle score—the score of such a kind that half the scores are better and half are not so good. It is a measure easy to understand and easy to calculate. It is for many purposes a more reliable measure of central tendency than the 'average', because it is not markedly affected, as is the average, by a few extreme cases.

To find the median, one counts in from either end of the ordered distribution of scores until he finds the middle score. If one has sorted the test papers according to size of score, then the median may be taken as the score on the middle paper. (If there are an even number of papers, the score on the two middle papers may be averaged together for the median). If one has written down the scores as shown in Tables II and III, he can count in by the same method from one end of the table. For example, in Table II there are 39 scores, and the middle paper is therefore the 20th, for there are 19 papers better and 19 not so good. The 20th paper is one of the 8 which received scores of 53. The approximate median score is therefore 53.

In more precise work it may be desirable to take the median, not at any particular score, but as the point on the score-line above and below which there is an equal number of measures. We can then express the median to the nearest tenth or hundredth of an integer. This may be done, in the example taken, as follows: Counting in from the bottom of the table, as before, we have $1 + 1 + 3 + 4 + 4 + 5 = 18$. $39 \div 2 = 19.5$. We now take the median score as that numbered 19.5, or exactly half the number of scores. As we wish to find the theoretical score No. 19.5, and as in our counting we have passed 18 scores,

it is plain that we must go 1.5 scores farther (19.5 — 18 = 1.5). We may assume that the 8 papers scored 53 would, if they were more accurately scored, include scores from 53.0 up to 53.9, and that they may be thought of as spread *evenly* over this interval. The median paper, for which we are looking, would then be $\frac{1.5}{8.0}$, or $^{15}\!/_{80}$ of the distance across the interval. The distance across the interval from 53 to 54 is one unit, and $^{15}\!/_{80}$ of 1 is .18, or approximately .2. This is the amount by which the median score exceeds 53. Adding it to 53 gives 53.2, the true median.

In calculating the median in an arrangement of scores such as that of Table III, the scheme of "interpolation" just explained is always necessary, and is here perhaps easier to understand. In Table III there are 80 scores represented (sum of frequency column). $\frac{80}{2} = 40$. $1 + 2 + 4 + 6 + 8 + 11 = 32$. $40 - 32 = 8$. $^{8}\!/_{12}$ of $10 = 6.7$. $80 + 6.7 = 86.7$, the median. It is here seen that the 12 scores which include the one sought, the 40th, are spread between 80 and 89. If they may be assumed to be spread evenly across this interval, then the 40th is eight-twelfths of the distance across. The distance across the interval is 10 units. Adding eight-twelfths of 10, or 6.7 to the beginning of the interval, or 80, will then locate the median. In making this interpolation, one should be especially careful that the correction (here 6.7) is added to the right number. Many more mistakes are made in this step than in finding the correction. It is a good plan to check the work by counting from the other end of the table. In Table III this would be done as follows: $2 + 4 + 7 + 6 + 7 + 10 = 36$. $40 - 36 = 4$. $^{4}\!/_{12}$ of $10 = 3.3$. $90 - 3.3 = 86.7$, the same result. In counting down in this way from the large scores to the small ones, one must remember that the correction (3.3) is in this case to be subtracted from the upper end of the interval,

not added. The upper end of the interval 80-89 is taken as 90 because 89 is only an abbreviation for 89.99999 +, which goes as close to 90 as you please.

It should be pointed out that in a list of scores such as that of Table II the score 53 may mean either 53.0 to 53.9, or 52.1 to 53.0, or 52.5 to 53.4. That is, all papers may be called 53 if the exact score lies between 53 and 54, or if it lies between 52 and 53, or if their score is somewhere in the interval whose middle is 53. To make sure which of these three things the score 53 means, the tester should consult the account of the scoring plan which appears in the teacher's manual sent with the test. In practice, however, the score 53 may be assumed to mean 53 to 53.9 for all common scales except those of handwriting and English composition.

The calculation of the median should be thoroughly mastered. It is fundamental for all uses of standardized tests for comparing one group with another, or with their own past performance. The reader will do well to make up and work out examples similar to those shown here until the described checking plan shows him that his computations are reliable.

Summary of Steps in the Calculation of the Exact Median

1. Make a frequency table, collecting the scores if necessary into about ten or fifteen groups.
2. Find half the number of scores. Call this $\frac{N}{2}$.
3. Count in on the frequency column, from the smaller scores toward the larger, until adding another number would make the sum larger than $\frac{N}{2}$.
4. Subtract the sum so found from $\frac{N}{2}$.
5. Divide the remainder by the number of scores in the next group on the frequency line.
6. Multiply the result by the number of units in the interval on the score line. (If the scores are not grouped, the interval on the score line is 1).

7. Add this number to the beginning of the interval opposite the number used in step 5.

8. This is the exact median.

9. Check by counting in from the opposite end of the table, in this case subtracting the correction (found in step 6) from the upper end of the middle interval instead of adding it to the lower end.

The Quartile Deviation

Besides a measure of the "central tendency" of his class, the tester will often want a good measure of their variability, or range of score. The total range, found by subtracting the lowest score from the highest one, is not very informing, because it is too much affected by one or two extreme scores. A single very bright or very dull pupil may give one an entirely wrong idea of the amount of variability characteristic of his class. It would be better if a few of the extreme scores could be disregarded. The common practice in this matter is to disregard the best quarter and the poorest quarter of the scores. Subtracting the lowest from the highest of the scores which then remain will give a measurement of the amount of variability in the *middle half* of the class. Dividing this amount by two will then show how far the most extreme pupils in the middle half of the class deviate from the median. This could be worked out directly from the stack of test papers arranged according to size of score, as was explained for the median. It could also be worked out from tabular arrangements of scores, such as those of Table II and Table III. The calculation for Table III would be as follows: $80 \div 4 = 20$. $1 + 2 + 4 + 6 = 13$. $20 - 13 = 7$. $\frac{7}{8}$ of $10 = 8.7$. $60 + 8.7 = 68.7$, the lowest quartile point. Similarly, counting in one-fourth of the way from the upper end of the scale we have $2 + 4 + 7 + 6 = 19$. $20 - 19 = 1$. $\frac{1}{7}$ of $10 = 1.4$. $110 - 1.4 = 108.6$, the upper quartile point. $108.6 - 68.7 = 39.9$, the range of

the middle half of the class. $39.9 \div 2 = 19.9$, the quartile deviation, or the amount that the most extreme persons in the middle half of the class deviate from the median.

The Mean Deviation

A slightly different measure of variability, the mean or average deviation, is sometimes preferred because it takes into account all the pupils of the class. It may be calculated simply by subtracting from the median each of the scores smaller than the median, and subtracting the median from each of the scores larger than itself. This, of course, shows how far each pupil deviates from the median. Adding these deviations and dividing by the number of scores then gives the average or mean deviation.

The Standard Deviation

Still another measure of variability, the standard deviation, is computed in a similar way, except that the deviations are each squared. The sum of the squares is found and divided by the number of scores, and the square root taken of the result. This is the measure of variability most commonly used in statistical work which is carried out on a large scale. It is the expression for deviation which will ordinarily be found in printed accounts of experiments with educational measurements.

The Measurement of Relationship

It is often desirable to measure the closeness of relationship between two sets of scores. For example, one may wish to know how closely two intelligence tests agree with each other in their ratings of a certain group of children, or how closely arithmetic and reading scores agree with each other (whether pupils scoring high in arithmetic are likely to score high in reading), or how closely ability in oral reading corresponds with ability in silent reading. The amount of relationship present can be shown either graphically or numerically.

Graphic Methods. One of the commonest graphic devices is the scatter diagram. This is made by laying off at equal distances along one side of a piece of cross-section paper the scores in one test, and laying off at right angles the scores in the other test. A dot may then be made on the cross-section paper where the lines representing the two scores of a given pupil intersect. Figure I

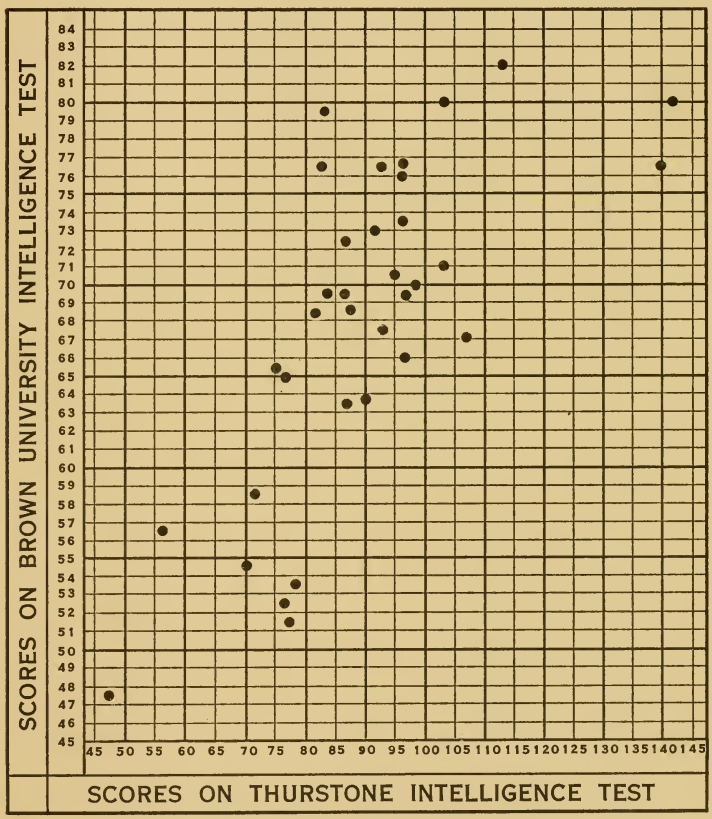

FIGURE I.

shows the general plan. A certain pupil had a score of 71 on the Thurstone test and a score of 58.8 on the Brown University test. Find the dot which represents him. Another pupil had a score of 107 on the Thurstone test and a score of 67.1 on the Brown University test. Find his location on the diagram.

If, as shown on this figure, the scores laid out horizontally are placed on the diagram so as to increase from left to right, and those laid out vertically are arranged to increase from the bottom of the sheet toward the top, then pupils who get high scores on both tests will be located in the upper right-hand corner of the figure, those who get low scores on both tests in the lower left-hand corner, and those standing medium in both tests will fall near the center of the figure. This means that, if the two sets of scores are plotted so as to be equally spaced along the two sides of the paper, the closeness of the relationship between them can be seen from the closeness with which the dots are grouped about a line running from the lower left-hand corner to the upper right-hand corner of the figure. This is what we wish to read from the figure.

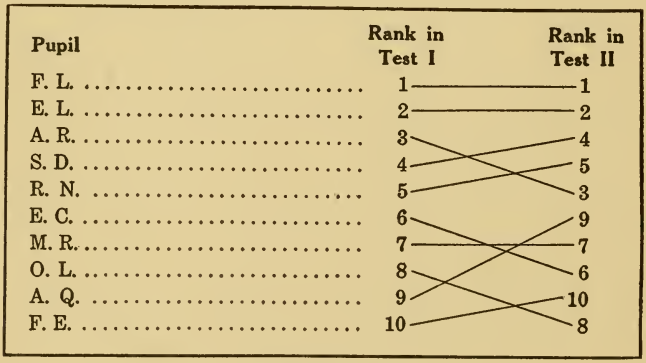

FIGURE II.

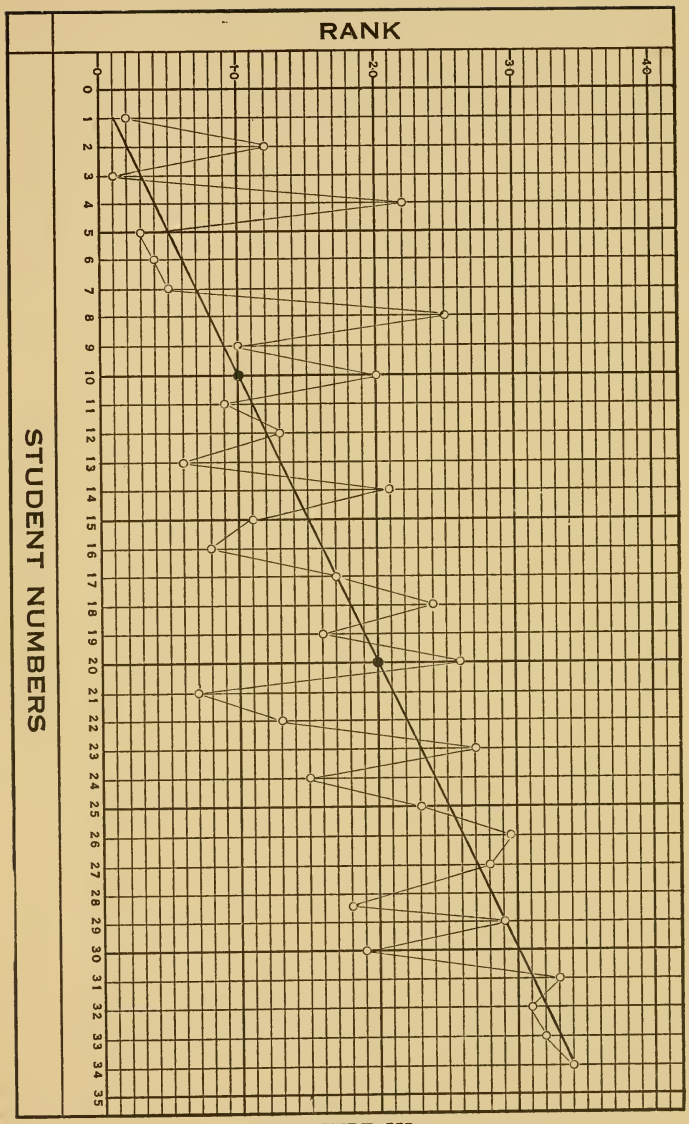
RANK
0
10
20
30
40
STUDENT NUMBERS
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35

FIGURE III.

Another simple graphic device is made by arranging the pupils' names in order according to their standing on one test, and then showing opposite each name the rank of the pupil in each test. If lines are drawn from No. 1 in one column to No. 1 in the other, from No. 2 in one column to No. 2 in the other, and so on, as shown in Figure II, the amount of crossing of the lines will show the amount of disagreement between the two tests, and the steepness of certain lines will show the extreme degrees of disagreement, or the amount certain pupils change in rank from one test to the other.

Another simple graph can be constructed by plotting the ranks of one over the other, as in Figure III. Here the pupils have been arranged in order according to their scores on the Thurstone intelligence test, and the ranks plotted as the straight line. (The names or key-numbers of the pupils are written along the bottom of the sheet). The ranks in the Brown University intelligence test are plotted to the same base for comparison. If all the pupils should rank exactly the same in both tests, the two lines would coincide. The amount of disagreement between the two tests is therefore shown by the amount in which the lines diverge from each other.

A graph of a kind commonly seen in newspapers and magazines is shown in Figure IV. Different kinds of shading are used to represent different degrees of the quality represented. For example, the bars might represent I. Q.'s of a class according to each of the two intelligence tests: the heavily hatched portion of each bar might represent I. Q.'s between 70 and 79, the next lighter por-

FIGURE IV.

tion I. Q.'s between 80 and 89, and so on. Whether the two intelligence tests agree in their distribution of various degrees of ability among the members of a class could then be seen at a glance.

Still another common method of representing relationship is the use of parallel bars whose total length depends on the amount of the thing represented. Two bars may be made for each pupil, one representing, say, his rank in class according to one test, and the other his rank according to the other test. These bars may be differently shaded or differently colored. The degree in which the various pairs of bars agree in length would then show the amount of agreement between the two tests. It is best to arrange the names of the pupils (written below the bars) in order according to one measure or the other. This makes the figure considerably easier to read.

Graphic devices give a striking representation of results, but not one which easily allows precise statements as to the amount of relationship discovered. One cannot always say, from an examination of graphs, whether the agreement between a certain pair of tests is greater or less than the agreement between another pair; and he can never say *how much* greater or less it is. For such information we must rely upon numerical rather than graphic treatment of the data.

Numerical Methods: Frequency Table. The numerical methods of indicating the degree of relationship between two sets of scores indicate the facts very definitely. One of the best is the frequency table in terms of changes of rank. If all the pupils are ranked according to one test and then according to the other, and the differences in rank found, a frequency table can be made of the differences. The method is shown in Table IV. In Part II of this table it is seen that one pupil kept exactly the same rank in both tests, three pupils changed rank one place, five

pupils changed rank two places, and so on. Information of this kind is often of very high practical value.

Table IV, Part I

Student	Rank in Test I	Rank in Test II	Difference in Rank
A. C.	1	3	2
B. D.	2	1	1
E. J.	3	4	1
R. L.	4	2	2
S. M.	5	9	4
E. F.	6	6	0
R. S.	7	16	9
A. V.	8	5	3
S. Z.	9	7	2
L. M.	10	18	8
N. O.	11	10	1
P. Q.	12	17	5
N. M.	13	8	5
N. L.	14	11	3
V. W.	15	12	3
X. Z.	16	14	2
S. P.	17	15	2
S. D.	18	13	5

Table IV, Part II

Difference in Rank	Frequency
0	1
1	3
2	5
3	3
4	1
5	3
6	0
7	0
8	1

Sum of Differences in Rank. Another good method of measuring relationship between two sets of scores is to find the sum of the differences in the ranks given the pupils by the two tests. The sum of the right-hand column in Table IV, Part I, can be compared either with the sum of the differences in rank if the differences were made as large as possible, or with the sum in case the rankings were left to chance. The sum of the differences in rank when rank is left to chance is $\frac{n^2-1}{3}$, where n is the number of pupils. The sum of the maximum differences in rank can be found very quickly, even for a large group, if one observes that the column of differences in this problem takes the form of two ordered series, 1, 3, 5, 7, 9, etc. (see Table V), and that they can therefore be added by the formulae, $l = a + (n-1)d$, and $s = (a+l)\frac{n}{2}$, in which a is the first term of the series, l the last term, n the number of terms, d the common difference between the terms, and s the sum of the terms. Comparing the sum of the actual differences in rank with the sum of the maximum differences, and with the sum of the differences when rankings are left to chance, gives a fairly good idea of the amount of agreement between two tests.

Table V

Rank in Test I	Rank in Test II	Differences in Rank
1	10	9
2	9	7
3	8	5
4	7	3
5	6	1
6	5	1
7	4	3
8	3	5
9	2	7
10	1	9

$l = a + (n - 1)d = 1 + (5 - 1)2 = 9$
$s = (a + l)\frac{n}{2} = (1 + 9)\frac{5}{2} = 25$
$\Sigma = 25 \times 2 = 50$

The Coefficient of Correlation. The measure for relationship which is most commonly used is the coefficient of correlation. This is a much more compact way of representing results, as the whole story is told by a single number, or at most by a pair of numbers. The devices described above express the relationship either by means of space-filling tables or by some rather long phrase, such as "the average change of rank is 12 out of a possible 38," or "the sum of the differences in rank is 1,283, as compared with a random-change sum of 1,822, and a maximum-change sum of 2,346." The coefficient of correlation says all this by one number, for example, .72. The formula leading to this convenient result takes several forms, but the one best meeting the needs of a teacher who is working with a single class may be expressed, with a slight change from its traditional statement, thus: $r = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)}$ where *r* stands for the coefficient of correlation, Σ for "the sum of," D for the differences in rank, and N for the number of cases (ordinarily the number of pupils who took both tests). An illustration will make the use of the formula clear:

Table VI

Student	Rank in Test I	Rank in Test II	Difference in Rank	Difference Squared
A. C.	1	3	2	4
B. D.	2	1	1	1
E. J.	3	4	1	1
R. L.	4	2	2	4
S. M.	5	9	4	16
E. F.	6	6	0	0
R. S.	7	16	9	81
A. V.	8	5	3	9

Student	Rank in Test I	Rank in Test II	Difference in Rank	Difference Squared
S. Z.	9	7	2	4
L. M.	10	18	8	64
N. O.	11	10	1	1
P. Q.	12	17	5	25
N. M.	13	8	5	25
N. L.	14	11	3	9
V. W.	15	12	3	9
X. Z.	16	14	2	4
S. P.	17	15	2	4
S. D.	18	13	5	25
R. O.	19	20	1	1
L. K.	20	19	1	1
				288

$$r = 1 - \frac{6 \Sigma D^2}{N(N^2-1)} = 1 - \frac{6 \times 288}{20(20^2-1)} = .78$$

$$PE = \frac{.6745(1-r^2)}{\sqrt{N}} = \frac{.6745(1-.78^2)}{\sqrt{20}} = .06$$

Summary of Steps in Rank-Difference Method of Calculating the Coefficient of Correlation

1. Find the rank of each pupil in each of the two tests, and place these numbers opposite his name, as shown in Table VI.
2. Subtract each pupil's rank in one test from his rank in the other, obtaining column headed "Difference in Rank."
3. Square each of these differences, obtaining column headed "Difference Squared."
4. Add this column of squares, obtaining Σ.
5. Multiply this sum by 6.
6. Square the number of pupils, subtract 1, and multiply the result by the number of pupils.
7. Divide the result of step 5 by the result of step 6.
8. Subtract the quotient from 1.
9. This is the coefficient of correlation.

Accuracy of the Coefficient. The accuracy of a correlation coefficient is proportional to the number of scores involved and to the closeness of the agreement between them. The formula for the "probable error" of a coefficient of correlation is, $PE = \frac{.6745}{\sqrt{N}}(1 - r^2)$, where N is the number of pupils. A coefficient of correlation of .78 with a probable error of .06 would mean that the coefficient may be taken to lie between .72 and .84. The probable error increases rapidly as the number of cases falls off, and coefficients of correlation should not be calculated for classes smaller than twenty.

Meaning of the Size of the Coefficient. A perfect agreement between two sets of scores—the same rank given each pupil by both tests—would give a coefficient of correlation of +1.00; perfect reversal of ranking would give —1.00; and the absence of any significant relationship between the two rankings would give 0. The significance of any obtained coefficient therefore depends upon how closely it approaches 1. In practice it is found that a coefficient of correlation less than .20 or .30 indicates an agreement so slight as to be insignificant; that a coefficient between .20 or .30 and .50 or .60 indicates a significant but not a very close agreement; that a coefficient between .50 or .60 and .70 or .80 indicates a close agreement; and that a coefficient higher than this indicates an agreement that is very close, indeed. The exact meaning of such a result, however, depends upon the kind of data from which it was derived, and the teacher should be very careful in interpreting coefficients of correlation. For example, small consistent changes of rank throughout the class may make the coefficient of correlation quite low, while as a matter of fact the agreement between the two sets of scores may be close enough for the purpose the teacher has in mind. This would be true if the purpose was to classify pupils into groups of uniform ability, for many small changes of rank would not here be significant,

since they would not cause pupils to move *far* up or down the scale and thus different tests would not place them in different sections.

The Correlation Table. Perhaps for the classroom teacher the most useful measure of relationship, in addition to the frequency table described on pages 87 and 88, is the correlation table. This is very similar in construction to the scatter diagram. An example will make it plain (Table VII):

Table VII

Scores in Otis Test	Scores in Illinois General Intelligence Examination					
	Below 68	68-81	82-95	96-109	110+	Totals
135+	..	1	2	4	3	10
115-134	..	4	9	8	7	28
95-114	2	11	21	4	3	41
75-94	7	15	10	2	..	34
Below 75	3	3	1	..	..	7
Totals	12	34	43	18	13	120

This table reads as follows: One pupil received a score above 135 on the Otis test, but a score between 68 and 81 on the Illinois Examination—was placed in the upper fifth or quintile of his class by one test and in the fourth quintile by the other; two pupils received a score above 135 on the Otis test, but a score between 82 and 95 on the Illinois Examination—were placed in the upper quintile by one test and in the third quintile by the other; and so on. The pupils about whom the tests agree lie on the diagonal running from lower-left to upper-right. Adding up these numbers, we have $3 + 15 + 21 + 8 + 3 = 50$,

which is 42% of the total number, 120. The pupils about whom there is substantial but not complete agreement will lie adjacent to the diagonal. Adding up these numbers, we have $7 + 11 + 9 + 4 + 7 + 4 + 10 + 3 = 55$, which is 46% of 120. Combining 42% and 46% we have 88% as the portion of the class in which the two tests are in at least substantial agreement. This is just the kind of information the teacher needs in practical work with intelligence measurements.

Another use for such a table is to show whether the data are of such a kind as to make the calculation of the coefficient of correlation worth while. If the larger number of the scores lie in an approximately straight line, or if a line drawn through the centers of each column so as to have an equal number of scores above and below it is straight, then the coefficient of correlation is applicable; but if such a line is curved, the coefficient of correlation is not applicable. It is therefore well to make such a table even if the correlation coefficient is to be computed by the method shown above, which does not directly utilize the table.

In making a correlation table one should make the spaced intervals (in this table, 68-81, etc.) equal to each other in each set of scores (though not necessarily in both sets), and should so select them that the line of totals at the bottom and sides of the table will at least roughly correspond to the "normal" distribution, that is, will take something like these percentage distributions: 7%, 24%, 38%, 24%, 7%. This can be done by counting in from either end of the ordered distribution of scores (as in Table II or Table III), until one has passed about 7% of the cases. The two scores thus located can then be subtracted, and the difference divided by three (if quintiles are to be used); this will give the approximate size of the score interval. When very large numbers have been tested, a finer distribution can be made, if desired, by

dividing the scale into sevenths instead of fifths. In this case the result of the subtraction, just described, will be divided by five to get the size of the score interval.

Summary. The three types of measures which the teacher will need are: measures of central tendency, measures of variability, and measures of relationship. The best measures of central tendency are the average (technically called the mean), and the median, and of these two the median is almost universally used in work with standardized tests. The best measures of variability are the quartile deviation, the average or mean deviation, and the standard deviation. The best measures of relationship, in addition to various types of graphs, are the tables of frequency of difference in rank, the sum of the differences in rank, the coefficient of correlation, and the correlation table. The coefficient of correlation is the measure of relationship commonly used in statistical work, but the frequency table and the correlation table will probably be more useful in work done with standardized tests by the classroom teacher.

INDEX